Seeing the Unseen: Composing Outliers for Compositional Zero-Shot Learning

Chenchen Jing[1,2], **Mingyu Liu**[3], **Hao Chen**[3], **Yuling Xi**[3], **Xingyuan Bu**[4], **Dong Gong**[5], **Chunhua Shen**[1,2*]
[1]College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou, China
[2]Zhejiang Key Laboratory of Visual Information Intelligent Processing, Hangzhou, China
[3]Zhejiang University, China
[4]Alibaba Group
[5]The University of New South Wales
{jingchenchen,chhshen}@zjut.edu.cn {mingyuliu, haochen.cad,xiyuling}@zju.edu.cn,
xingyuanbu@gmail.com, dong.gong@unsw.edu.au

Abstract

Compositional zero-shot learning (CZSL) is to recognize unseen attribute-object compositions by learning from seen compositions. The distribution shift between unseen compositions and seen compositions poses challenges to CZSL models, especially when test images are mixed with both seen and unseen compositions. The challenge will be addressed more easily if a model can distinguish unseen/seen compositions and treat them with specific recognition strategies. However, identifying images with unseen compositions is non-trivial, considering that unseen compositions are absent in training and usually contain only subtle differences from seen compositions. In this paper, we propose a novel compositional zero-shot learning method called COMO, which **com**poses **o**utliers in training for distinguishing seen and unseen compositions and further applying specific strategies for them. Specifically, we compose attribute-object representations for unseen compositions based on primitive representations of training images as outliers to enable the model to identify unseen compositions in inference. At test time, the method distinguishes images containing seen/unseen compositions and uses different weights for composition classification and primitive classification to recognize seen/unseen compositions. Experimental results on three datasets show the effectiveness of our method in both the closed-world setting and the open-world setting.

1 Introduction

Compositional generalization, understanding unseen combinations composed of seen primitives, is one of the fundamental properties of human intelligence [Fodor and Pylyshyn, 1988]. To evaluate such ability of vision models, compositional zero-shot learning (CZSL) [Misra *et al.*, 2017; Purushwalkam *et al.*, 2019] requires recognizing unseen attribute-object compositions by learning from seen compositions in training images, as shown in Fig. 1 (a). In particular, CZSL usually follows a practical and challenging generalized zero-shot learning setting [Pourpanah *et al.*, 2022; Chen *et al.*, 2020; Liu *et al.*, 2021a], where the test set contains both seen and unseen classes, in evaluation. Since the label combinations of unseen compositions were never observed during training, the distribution shift between the seen compositions and unseen compositions is significantly large, which poses challenges to CZSL models [Atzmon *et al.*, 2021; Mancini *et al.*, 2022]. The learned correlations of models in training may be detrimental at test time.

Previous works deal with the distribution shift via invariant representation learning [Atzmon *et al.*, 2021; Zhang *et*

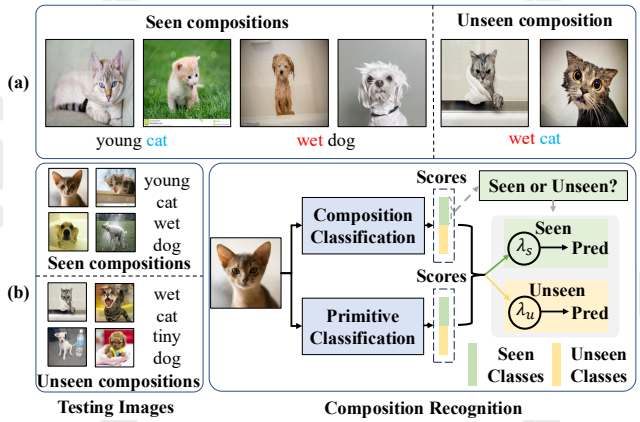

Figure 1: Illustration of identifying unseen compositions for compositional zero-shot learning. (a) shows two seen compositions in the training set of the MIT-States [Isola *et al.*, 2015] on the left and an unseen composition in the testing set on the right. (b) shows that the testing images contain both seen and unseen compositions, and our method first identifies whether testing images contain seen compositions and then uses different strategies for seen/unseen compositions. The λ_s/λ_u denotes the weights for combining composition classification and primitive classification for seen/unseen compositions, respectively.

*corresponding author

al., 2022], prototypical representation learning [Ruis *et al.*, 2021] or propagating information of seen primitives for unseen compositions [Naeem *et al.*, 2021; Mancini *et al.*, 2022]. They aim to recognize unseen compositions well by adding structural constraints, which may reduce the useful correlations for seen compositions and influence the performance. However, considering the potential differences between unseen and seen compositions, the tasks can be handled more easily if a model can distinguish the seen and unseen compositions and treat them with specific recognition strategies, such as specific expert classifiers. In this work, we explore an explicit way to handle the distribution shift caused by unseen cases, separating images of seen compositions from those of unseen compositions and using different strategies for them. Nonetheless, identifying unseen compositions at test time is non-trivial. The unseen compositions are absent in model learning, and thus their distribution is unknown. Sharing similar primitives with seen compositions, they may be only subtly different from seen compositions.

To address these challenges, we present a novel compositional zero-shot learning method called COMO that **com**poses **o**utliers in training for identifying unseen compositions and further using different strategies for seen/unseen attribute-object compositions. The method composes image-level representations for unseen attribute-object compositions based on learned representations of seen primitives. These composed representations are regarded as outliers of the training distribution to encourage the model to identify out-of-distribution (OOD) samples at testing. Besides, the disentanglement of the primitive representations and the discriminability of the composed representations is considered to guarantee appropriate composed representations.

Specifically, our method combines a *composition classification* module and a *primitive classification* module for composition recognition, as shown in Fig. 1 (b). The former module characterizes compatibility between images and candidate compositions, while the latter module independently characterizes the compatibilities between images and the candidate attributes and objects. By enforcing the composition classification module to output the uniform distribution for composed representations, the module learns heuristics to identify images with unseen composition. Considering the two classification modules focus on different perspectives of an image and can supplement each other [Yang *et al.*, 2022; Wang *et al.*, 2023b; Huang *et al.*, 2024], we assign different weights for seen/unseen compositions to combine the two modules to achieve composition recognition. The experimental results on three widely used CZSL datasets under both the closed-world setting and the open-world setting show the effectiveness of the proposed method.

The contributions of this paper are summarized as:

1. We propose a novel compositional zero-shot learning method that can identify images with unseen compositions at test time and uses different strategies for seen/unseen compositions.
2. Our method composes unseen attribute-object compositions based on primitive representations to obtain outliers for the model to identify unseen compositions.

2 Related Work

Compositional zero-shot learning. The task of compositional zero-shot learning aims to recognize unseen attribute-object compositions by learning from seen compositions. Existing methods mainly achieve the task via composition classification with a composition classifier [Misra *et al.*, 2017; Naeem *et al.*, 2021], or primitive classification, which independently recognize attributes and objects, [Li *et al.*, 2020; Purushwalkam *et al.*, 2019], or combines composition classification and primitive classification for better contextuality [Yang *et al.*, 2022; Wang *et al.*, 2023b]. With the recent advance in pre-trained vision-language models, CLIP-based CZSL methods [Nayak *et al.*, 2023; Lu *et al.*, 2023; Huang *et al.*, 2024; Bao *et al.*, 2023] achieved state-of-the-art performance. CSP [Nayak *et al.*, 2023] first uses the CLIP [Radford *et al.*, 2021] in CZSL. They replace the classes in textual prompts with trainable attributes and object tokens. Troika [Huang *et al.*, 2024] jointly models the vision-language alignments for the attribute, object, and composition using the CLIP. PLID [Bao *et al.*, 2024] leverages pre-trained large language models to enhance the compositionality of the softly prompted class embedding. The aforementioned work mainly focuses on parameter-efficient fine-tuning of CLIP. By contrast, our method focuses on identifying unseen compositions and designing different strategies for seen/unseen compositions and only uses CLIP as the backbone.

Distribution shift in CZSL. Atzmon *et al.*[Atzmon *et al.*, 2021] points out that the distribution shift between seen compositions and unseen compositions is a fundamental challenge for CZSL. They propose to ensure conditional independence between attribute and object representations via causal inference to handle this issue. Naeem *et al.*[Naeem *et al.*, 2021] and Mancini *et al.*[Mancini *et al.*, 2022] use graph convolutional networks to extract attribute-object representations and propagate information from seen compositions to unseen compositions to improve the generalization ability. Ruis *et al.*[Ruis *et al.*, 2021] learned compositional prototypes of novel attribute-object combinations that reflect the dependencies of the target distribution. Zhang *et al.*[Zhang *et al.*, 2022] treated CZSL as a domain generalization task, and proposed to learn attribute-invariant and object-invariant representations for CZSL. Li *et al.*[Li *et al.*, 2023] believe the influence of the distribution shift for the composition classification is larger, and only use a primitive classification module for composition recognition. Different from these works, we explicitly separate the seen class instances from those of the unseen classes, and independently classify seen and unseen class data samples via domain-specific strategies.

Generalized zero-shot learning (GZSL). GZSL is a typical setting of zero-shot learning (ZSL) to imitate the human capability of recognizing samples from both seen and unseen classes [Pourpanah *et al.*, 2022]. Previous work explores embedding-based [Min *et al.*, 2020; Hu *et al.*, 2023] or generative-based [Verma *et al.*, 2021; Liu *et al.*, 2021b] methods for GZSL. In this work, we borrow the idea of embedding-based GZSL methods to deal with the distribution shift issue of CZSL, and propose a simple but effective method tailored for CZSL.

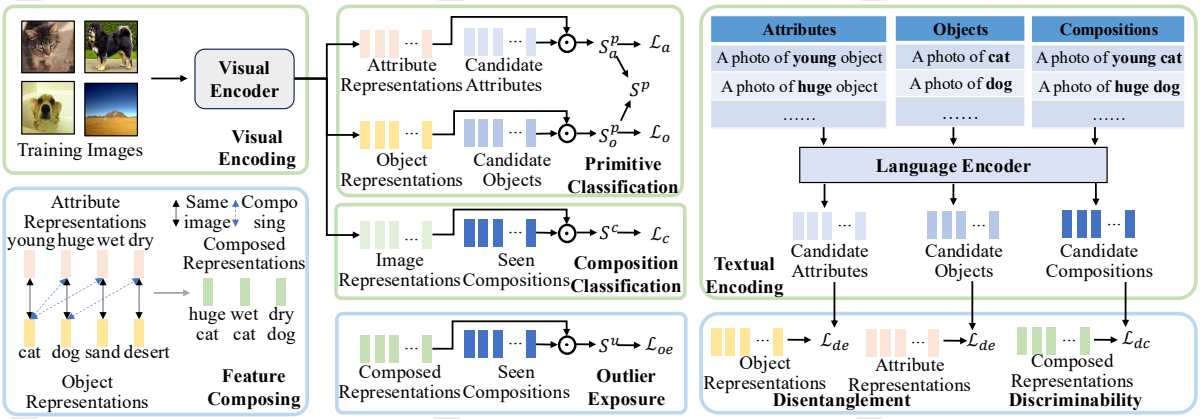

Figure 2: Overview of our method. In composition classification, the method characterizes primitive-level and composition-level compatibilities between the image and candidate compositions. The method composes representations for unseen compositions based on attribute and object representations of training images. An outlier exposure loss is used to enable identifying images with unseen compositions at testing. A disentanglement loss and a discriminability loss are used to guarantee the effectiveness of composed representations.

3 Method

In this section, we formulate the CZSL task and illustrate the proposed method, as shown in Fig. 2. Specifically, we illustrate how we achieve *composition recognition based on out-of-distribution (OOD) detection* by combining two classification modules, introduce how we *compose representations of unseen compositions as outliers* in training to enable the model to identify OOD samples, and introduce the *overall learning objective*.

3.1 Formulation

Compositional zero-shot learning aims at learning a model from limited compositions of attributes (*e.g.*, young, wet) and objects (*e.g.*, cat, dog) 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 $\mathcal{C}^s \cup \mathcal{C}^u \subset \mathcal{C}$. In particular, CZSL follows the generalized zero-shot learning setting [Pourpanah *et al.*, 2022; Chen *et al.*, 2020; Liu *et al.*, 2021a], where the training images only contain classes from the $\mathcal{C}^s$ and the images of testing set contains both seen classes and unseen classes.

Given a test image $I \in \mathcal{I}$, the CZSL task requires a model 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. Thus the test class set is only a subset of the compositional class set $\mathcal{C}$. By contrast, in the challenging open-world setting, the test class set is all possible compositions, *i.e.*, $\mathcal{C}^{test} = \mathcal{C}$. Formally, the model is required to model a score function $S : \mathcal{I} \times \mathcal{A} \times \mathcal{O} \rightarrow \mathbb{R}$ between an image I and a candidate composition. During inference, the candidate composition with the highest score is regarded as the final prediction.

3.2 Composition Recognition Based on OOD Detection

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 encoder of CLIP to obtain the representation of the [CLS] token $\boldsymbol{v} \in \mathbb{R}^d$, and then use three multi-layer perceptions (MLPs) to obtain image representation $\boldsymbol{v}^I \in \mathbb{R}^d$, attribute representation $\boldsymbol{v}^a \in \mathbb{R}^d$, and object representation $\boldsymbol{v}^o \in \mathbb{R}^d$, respectively.

For candidate compositions, attributes, and objects, we use the soft prompt [Nayak *et al.*, 2023] to obtain textual representations, by using a prompt template like "a photo of *[class]*". We feed the text encoder of CLIP with "a photo of *[attribute] [object]*", "a photo of *[attribute]* object", and "a photo of *[object]*" for each candidate composition, attribute and object to obtain their textual representations $\boldsymbol{T}^c \in \mathbb{R}^{N_c \times d}$, $\boldsymbol{T}^a \in \mathbb{R}^{N_a \times d}$, $\boldsymbol{T}^o \in \mathbb{R}^{N_o \times d}$, respectively. The N_c, N_a, and N_o are the numbers of candidate compositions, attributes, and objects, respectively. Specifically, the [attribute] and [object] tokens are trainable and initialized with the corresponding word embeddings of CLIP.

Two types of classification modules. To fully characterize the contextuality of attributes and objects, we use two types of classification modules: a *composition classification* module and a *primitive classification* module, to achieve composition recognition based on the above-mentioned representations. The composition classification models the composition compatibility $S^c(I, a, o)$. The primitive classification models the attribute compatibility $S^p_a(I, a)$, the object compatibility $S^p_o(I, o)$, and further obtain primitive-level scores $S^p(I, a, o)$. Specifically, for an input image I and a candidate composition $c_i = (a_j, o_k)$, the composition classification module measures the compatibility score as

$$\mathcal{S}^c(I, a_j, o_k) = \cos(\boldsymbol{v}^I, \boldsymbol{T}^c_i), \tag{1}$$

where $\cos(\cdot, \cdot)$ denotes the cosine similarity function of two vectors. The primitive classification module independently measures the compatibility scores between an image and the

attribute a_j and the object o_k, and computes the primitive-level scores by using the aforementioned primitive-level visual and textual representations as

$$\begin{aligned}\mathcal{S}^p(I, a_j, o_k) &= \mathcal{S}^p_a(I, a_j) + \mathcal{S}^p_o(I, o_k) \\ &= \cos(\boldsymbol{v}^a, \boldsymbol{T}^a_j) + \cos(\boldsymbol{v}^o, \boldsymbol{T}^o_k).\end{aligned} \tag{2}$$

We optimize the two modules in training with the cross-entropy loss as

$$\begin{aligned}\mathcal{L}_{cls} &= \mathcal{L}_c + \mathcal{L}_a + \mathcal{L}_o, \\ \mathcal{L}_c &= -\log \frac{\exp(\mathcal{S}^c(I, c_{gt}))}{\sum_{k=1}^{|\mathcal{C}^s|} \exp(\mathcal{S}^c(I, c_k))}, \\ \mathcal{L}_a &= -\log \frac{\exp(\mathcal{S}^p_a(I, a_{gt}))}{\sum_{k=1}^{|\mathcal{A}|} \exp(\mathcal{S}^p_a(I, a_k))}, \\ \mathcal{L}_o &= -\log \frac{\exp(\mathcal{S}^p_o(I, o_{gt}))}{\sum_{k=1}^{|\mathcal{O}|} \exp(\mathcal{S}^p_o(I, o_k))},\end{aligned} \tag{3}$$

where $c_{gt} = (a_{gt}, o_{gt})$, are the ground truth composition, attribute, and object for the image I, respectively. We set equal weights for the two modules to let them supplement each other.

Combining classification modules for recognition. At test time, we identify images with unseen compositions through OOD detection by using the maximum logit value before the softmax layer [Jung *et al.*, 2021], and then assign different fusion weights for the above-mentioned classification modules for composition recognition.

For a testing image, we select the maximum value of the predicted distribution l^{max} of the composition classification module over the seen compositions. By performing threshold comparison with a pre-defined threshold T, we can obtain $\hat{y}$, which is the predicted label about whether the image contains a unseen composition. $\hat{y}$ is set as 1 if $l^{max} <= T$ and 0 otherwise. Then we compute the final compatibility score for the image with a candidate composition as

$$\mathcal{S}(I, a, o) = \begin{cases}\lambda_u \mathcal{S}^p(I, a, o) + (1 - \lambda_u)\mathcal{S}^c(I, a, o), \hat{y} = 1, \\ \lambda_s \mathcal{S}^p(I, a, o) + (1 - \lambda_s)\mathcal{S}^c(I, a, o), \hat{y} = 0,\end{cases}$$

where λ_s and λ_u are weights of fusing the two scores for images with seen compositions and unseen compositions, respectively. Finally, the candidate composition with the highest overall score is predicted.

Specifically, the validation set is used to obtain the threshold T and two weights. We first cast identifying unseen compositions in the validation set as a binary classification task based on the maximum probability, and select the best threshold T by varying its value. Then the best threshold is used to classify images in the validation set into two groups: images with seen compositions and images with unseen compositions. We further use grid search to find the best fusion weights λ_s and λ_u on the validation, to combine the two classification modules.

3.3 Composing Outliers in Training for Composition Recognition

To identify images with unseen compositions, we borrow the idea of outlier exposure [Hendrycks *et al.*, 2018] to encourage the model to learn effective heuristics for out-of-distribution detection, by exposing OOD examples to the model.

Composing unseen compositions. In real applications, it may be difficult to know the distribution of testing images one will encounter in advance. Fortunately, in CZSL, the set of possible compositions can be composed using seen classes. Thus by using the primitive representations of seen compositions, representations of unseen compositions can be easily obtained.

Intuitively, we can build the representation for an unseen composition *wet cat*, by using the attribute representation of an image about *wet ground* and the object representation of an image about *young cat*. Motivated by the famous mathematics of computational linguistics "King – Man + Woman = Queen", we simply add two primitive representations to obtain composition-level representation for an unseen composition $u_i = (a_j, o_k)$ in model learning as

$$\boldsymbol{v}^u_i = \hat{\boldsymbol{v}}^a_j + \hat{\boldsymbol{v}}^o_k, \tag{4}$$

where $\hat{\boldsymbol{v}}^a_j$/$\hat{\boldsymbol{v}}^o_k$ is the attribute/object representation of the image containing attribute a_j or object o_k, respectively. Specifically, we compose features for all possible compositions that do not appear in the training set, rather than the unseen compositions in the validation set or the testing set. The GloVe [Pennington *et al.*, 2014] is used to obtain the feasibility calibration of each possible composition for filtering out infeasible compositions. In model learning, we randomly select the attribute/object representations of the current training batch to obtain representation for a given unseen composition. Specifically, in model learning, we can obtain various unseen compositions based on a batch of N_{bs} training images. For simplicity, we also generate representations for N_{bs} unseen compositions, by randomly sampling primitives of the batch.

Constraints for feature composing. To obtain appropriate composed representations for unseen compositions, we enforce some constraints on the composed representations for unseen compositions and the primitive representations of training images, taking into consideration the disentanglement of the primitive representations, and the discriminability of the composed representations.

We first introduce a **disentanglement** loss to learn more disentangled primitive representations, considering the complex semantic entanglement of the attribute and the object in an image. 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

$$\mathcal{L}_{de} = \cos(\boldsymbol{v}^a, \boldsymbol{T}^o_{gt}) + \cos(\boldsymbol{v}^o, \boldsymbol{T}^a_{gt}), \tag{5}$$

where $\boldsymbol{T}^o_{gt}$ and $\boldsymbol{T}^a_{gt}$ are the textual representations for the ground-truth object label and ground-truth attribute label, respectively. Intuitively, the disentanglement loss can reduce the mutual information between attribute representations and object representations. Note that we directly optimize the similarities to make the model **cannot** predict the attribute label a_{gt} with the object representation $\boldsymbol{v}^o$, and vice versa. An alternative option is to use a negative cross-entropy classification loss, which will lead to the value of the loss becoming negative infinity quickly.

We further introduce a classification loss over all composed unseen compositions, to enforce the composed representations to be **discriminative**. For an unseen composition u_i, the loss is calculated as

$$\mathcal{L}_{dc} = -\log \frac{\exp(\mathcal{S}^u(v_i^u, u_i))}{\sum_{k=1}^{|\mathcal{C}^{s,u}|} \exp(\mathcal{S}^u(v_k^u, u_k))}, \quad \mathcal{S}^u(v_i^u, u_i) = \cos(\boldsymbol{v}_i^u, \boldsymbol{T}_{u_i}^{c,u}), \tag{6}$$

where, $|\mathcal{C}^{s,u}|$ is all composed unseen compositions, and $\boldsymbol{T}^{c,u}$ is the textual representations for these unseen compositions obtained via the textual encoder. Intuitively, we encourage the composed feature to be close to the textual presentations of unseen compositions at the semantic level.

Outlier exposure. To encourage the model to identify unseen compositions in testing, we introduce a loss as

$$\mathcal{L}_{oe} = \mathrm{CE}(\boldsymbol{P}_i^u, \boldsymbol{P}_{uni}), \quad \boldsymbol{P}_i^u = \mathrm{softmax}(f(\boldsymbol{v}_i^u, \boldsymbol{T}^c)), \tag{7}$$

where $\boldsymbol{P}_{uni}$ is a uniform distribution over all seen compositions, and $\boldsymbol{P}_i^u$ is the predicted distribution of the composition classification module. $\mathrm{CE}()$ is the cross-entropy function. $f()$ is to compute cosine similarities between a composed representation $\boldsymbol{v}_i^u$ and all seen compositions $\boldsymbol{T}^c$. $\mathcal{L}_{oe}$ prevents composition classification from outputting confident predictions for OOD samples. Thus unseen compositions can be detected in testing via threshold comparison.

3.4 Optimization

We sum up all the above-mentioned losses to compute the overall loss to supervise the model learning. The overall loss is calculated as

$$\mathcal{L} = \mathcal{L}_{cls} + \alpha_1 \mathcal{L}_{de} + \alpha_2 \mathcal{L}_{dc} + \alpha_3 \mathcal{L}_{oe}, \tag{8}$$

where α_1, α_2, and α_3 are hyper-parameters. We use a two-stage training strategy for the proposed method. In the first stage, all the losses are used to train a outlier detector. In the second stage, we use only the classification loss to obtain the classifier. The model architecture is the same for two stages.

4 Experiments

4.1 Experimental Settings

Datasets. We conduct experiments on three widely used datasets, UT-Zappos [Yu and Grauman, 2014], MIT-States [Isola *et al.*, 2015], and C-GQA [Naeem *et al.*, 2021]. UT-Zappos is a synthetic 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. MIT-States consists of $53,753$ crawled web images labeled with 1962 attribute-object. The dataset contains $1,262$ seen and 300/400 unseen compositions for training and validation/testing, respectively. C-GQA contains over $9,000$ common compositions and is split into $5,592$ seen and $1,040$/923 unseen compositions for training and validation/testing, respectively. Note that in validation/testing set of each dataset, the number of appeared seen compositions is equal to that of unseen compositions.

Metrics. We report the standard metrics of CZSL evaluation protocol in both closed-world and open-world settings, including the best seen accuracy (**S**), the best unseen accuracy (**U**), the best harmonic mean (**HM**) between the seen and unseen accuracy, and the area under the curve (**AUC**) of unseen versus seen accuracy. In the standard evaluation, to counteract biases of models for seen classes, the calibration bias, a scalar, is added to scores of unseen classes [Purushwalkam *et al.*, 2019]. By varying the value of the calibration bias, the best seen/unseen accuracy, the best HM, and the AUC can be computed. Specifically, the AUC is the area under the curve of obtained seen accuracies and unseen accuracies. Considering that the best seen/unseen accuracy, and the best HM is obtained with a specific calibration bias, the AUC is thus able to better describe the overall performance of a model. In the open-world setting, the GloVe [Pennington *et al.*, 2014] is used to filter out infeasible compositions.

Implementation details. We build our method with two backbones, the ResNet [He *et al.*, 2016] and CLIP [Radford *et al.*, 2021]. For the CLIP architecture, ViT-L/14 is used as the previous work [Lu *et al.*, 2023]. For the ResNet backbone, we use a frozen ResNet-18 to obtain visual representations as [Yang *et al.*, 2022]. We report the results of our method with both backbones on three datasets under two settings (Sec. 4.2). For other analyses, we use the CLIP backbone because the performance with it is superior. In outlier composing, we filter infeasible pairs by removing pairs with a Glove score lower than 0.5. In the first stage, we perform the grid search with the validation set to find the threshold to identify unseen compositions. In the second stage, we perform the grid search on the validation set by varying λ_u in {0.0,0.1,0.2,...1.0} and λ_s in {0.0,0.1,0.2,...1.0}. For the CLIP backbone, the training epochs for each dataset as 5/15 for the two stages, respectively. In the first stage, the hyper-parameters α_1, α_2, and α_3 are set as (0.1, 0.1, 5.0) for the UT-Zappos, (0.01, 0.01, 1.0) for MIT-State, and (0.1, 0.5, 1.0) for C-GQA, respectively. For the ResNet backbone, the training epochs for each dataset as 50/100 for the two stages, respectively. The hyper-parameters α_1, α_2, and α_3 are set as (5.0, 0.1, 5.0) for the UT-Zappos, (5.0, 0.1, 1.0) for MIT-States, and (0.1, 1.0, 1.0) for C-GQA, respectively.

4.2 Main Results

We compare our method with various state-of-the-art methods, including both methods without CLIP and CLIP-based methods. All methods use a frozen visual feature extractor. Results of all methods on the test split of MIT-States [Isola *et al.*, 2015], UT-Zappos [Yu and Grauman, 2014], and C-GQA [Naeem *et al.*, 2021] under the standard closed-world setting are listed in Tab. 1. We observe that our method outperforms all methods in terms of AUC, which comprehensively evaluate the performance of CZSL models. The main reason is that our method can identify unseen compositions and separately classify seen and unseen compositions to alleviate the influence of distribution shift. Thus our method can achieve satisfactory performance for seen compositions and unseen compositions in testing with different values of the calibration bias. Note that we didn't compare with Troika [Huang *et al.*, 2024], CAILA [Zheng *et al.*, 2023], and RAPR [Jing

	Closed-world Model	Venue	MIT-States AUC	HM	S	U	C-GQA AUC	HM	S	U	UT-Zappos AUC	HM	S	U
Without CLIP	CGE$_{ff}$ [Naeem *et al.*, 2021]	CVPR'21	5.1	17.2	28.7	25.3	2.5	11.9	27.5	11.7	26.4	41.2	56.8	63.6
	SCEN [Li *et al.*, 2022]	CVPR'22	5.3	18.4	29.9	25.2	2.9	12.4	28.9	12.1	32.0	47.8	63.5	63.1
	DECA [Yang *et al.*, 2022]	TMM'22	5.3	18.2	29.8	25.2	-	-	-	-	31.6	46.3	62.7	63.1
	CANet [Wang *et al.*, 2023b]	CVPR'23	5.4	17.9	29.0	26.2	3.4	14.5	30.0	13.2	33.1	47.3	61.0	66.3
	COMO		5.5	18.0	28.4	26.3	3.6	14.8	29.7	14.9	33.5	49.0	60.3	65.3
With CLIP	CSP [Nayak *et al.*, 2023]	ICLR'23	19.4	36.3	46.6	49.9	6.2	20.5	28.8	26.8	33.0	46.6	64.2	66.2
	HPL [Wang *et al.*, 2023a]	IJCAI'23	20.2	37.3	47.5	50.6	7.2	22.4	30.8	28.4	35.0	48.2	63.0	68.8
	DFSP [Lu *et al.*, 2023]	CVPR'23	20.6	37.3	46.9	52.0	10.5	27.1	38.2	32.9	36.9	47.2	66.7	71.7
	DLM [Hu and Wang, 2024]	AAAI'24	20.0	37.4	46.3	49.8	7.3	21.9	32.4	28.5	39.6	52.0	67.1	72.5
	ProLT [Jiang and Zhang, 2024]	AAAI'24	21.1	38.2	49.1	51.0	11.0	27.7	39.5	32.9	36.1	49.4	66.0	70.1
	PLID [Bao *et al.*, 2024]	ECCV'24	22.1	39.0	49.7	52.4	11.0	**27.9**	38.8	33.0	38.7	52.4	67.3	68.8
	COMO		**22.4**	**39.3**	**50.4**	**52.7**	**11.4**	27.7	**40.3**	**33.3**	**43.7**	**56.6**	**68.5**	**74.0**

Table 1: The results on CZSL datasets in the *closed-world* setting. All methods use a frozen visual feature extractor. The best results of methods without CLIP are underlined and the best results of methods with CLIP are bold.

	Open-world Model	Venue	MIT-States AUC	HM	S	U	C-GQA AUC	HM	S	U	UT-Zappos AUC	HM	S	U
Without CLIP	CGE$_{ff}$ [Naeem *et al.*, 2021]	CVPR'21	0.7	4.9	29.6	4.0	0.30	2.2	28.3	1.3	21.5	39.0	58.8	46.5
	KG-SP$_{ff}$ [Karthik *et al.*, 2022]	CVPR'22	1.0	6.7	23.4	7.0	0.44	3.4	26.6	2.1	22.9	39.1	58.0	47.2
	Co-CGE$_{ff}^{CW}$ [Mancini *et al.*, 2022]	TPAMI'22	1.1	6.5	28.2	6.0	0.29	2.1	28.9	1.2	20.1	36.1	59.5	41.5
	COMO		1.4	7.9	26.7	7.9	0.48	3.5	29.4	2.4	23.7	40.8	60.3	48.2
With CLIP	CSP [Nayak *et al.*, 2023]	ICLR'23	5.7	17.4	46.3	15.7	1.2	6.9	28.7	5.2	22.7	38.9	64.1	44.1
	HPL [Wang *et al.*, 2023a]	IJCAI'23	6.9	19.8	46.4	18.9	1.5	7.5	30.1	5.8	24.6	40.2	63.4	48.1
	DFSP [Lu *et al.*, 2023]	CVPR'23	6.8	19.3	47.5	18.5	2.4	10.4	38.3	7.2	30.3	44.0	66.8	60.0
	PILD [Bao *et al.*, 2024]	ECCV'24	7.3	20.4	49.1	18.7	2.5	10.6	39.1	7.5	30.8	46.6	67.6	55.5
	COMO		**7.7**	**20.8**	**49.2**	**19.5**	**3.1**	**12.4**	**40.7**	**8.8**	**33.8**	**50.1**	**65.0**	**61.2**

Table 2: The results on CZSL datasets in the *open-world* setting. All methods use a frozen visual feature extractor. The best results of methods without CLIP are underlined and the best results of methods with CLIP are bold.

et al., 2024], because they use adapter or LoRA [Hu *et al.*, 2021] module in the transformer-based image encoder.

Tab. 2 shows the results on three datasets in the open-world setting. The proposed method outperforms all other methods, which demonstrates the effectiveness of our method for open-world compositional zero-shot learning. We observe that the performance gap between other methods and our method in the open-world setting is larger than that in the closed-world setting. A possible reason is in feature composing, we compose all possible compositions based on training images, which aligns with the open-world setting, where all possible compositions should be considered. Note that we use identical model weights, threshold, and fusion weights λ_s and λ_u for the two settings.

Model	AUC	HM	S	U
Baseline	41.9	55.4	68.0	72.1
Composition	38.2	51.1	65.2	72.6
Primitive	39.2	55.0	**68.8**	64.7
Real images	42.6	55.4	68.3	73.7
Generated images	42.8	55.9	68.0	74.1
sub	43.4	56.4	68.7	74.0
mul	43.4	56.0	68.7	73.4
cat	43.2	56.1	67.8	73.8
w/o composing	43.0	55.7	68.1	73.8
w/o $\mathcal{L}_{oe}$	42.7	55.9	68.2	74.1
w/o $\mathcal{L}_{de}$	43.4	56.6	68.3	74.1
w/o $\mathcal{L}_{dc}$	42.8	55.9	68.2	**74.2**
Oracle	48.3	65.2	68.7	73.2
COMO	**43.7**	**56.6**	68.5	74.0

Table 3: Ablations on UT-Zappos in closed-world setting.

4.3 Ablation Studies

To study the effectiveness of several key designs of our method, we evaluate different variants of our model on the UT-Zappos in the closed-world setting with CLIP.

Effect of identifying images with unseen compositions. We first investigate the effectiveness of identifying unseen compositions in testing and separately classifying images with unseen compositions and seen compositions. In Tab. 3, we provide the results of our baseline model with only the two classification modules. The AUC of the model is much lower than our full model. The results of the composition classification module and the primitive classification module of the baseline model are also reported. Formally, we set both λ_u and λ_s as 0 to obtain the model "Composition", and set them as 1 to obtain the model "Primitive". Both models are inferior to the baseline models. We can also obtain an oracle model by using the ground truth binary label for whether an image contains a seen composition. We observe that the model significantly outperforms the full model in terms of AUC, which shows the two classification modules are complementary to each other and the idea of detecting images of unseen pairs is promising. Note that the best seen/unseen accuracies of the Oracle model don't always outperform other models. The reason is that it also searches fusion weights to softly combine the classification modules, rather than directly use a specific module for seen or unseen compositions.

Effect of each component. Then we study the influence of

Dataset	Train	T	λ_u, λ_s	λ_u,λ_s (SGS)
UT-Zappos	320	0.1	5.4	0.5
MIT-States	825	0.1	150	30
CGQA	1530	0.1	610	70

Table 4: Time cost (s) for training one epoch, finding the best threshold and best fusion weights. The SGS denotes a simplified grid search with a smaller range.

composing unseen attribute-object compositions as outliers. We ablate the outlier composing to obtain the model "w/o composing", which also uses the MSP to identify images with unseen compositions. As shown in Tab. 3, the model performs slightly better than the baseline model, but is much worse than our full model, which demonstrates that the composed outliers bring substantial improvements. We further ablate the outlier exposure loss, the disentanglement loss, and the discriminability loss, to obtain three models, respectively. The comparisons demonstrate the three losses are also beneficial on both datasets.

Effect of composing primitive representations via addition. In our implementation, we add two primitive representations to obtain representation for an unseen composition. To investigate the effect of this design, we use subtraction and multiplication to obtain two models ("sub" and "mul"), respectively. We also concatenate the two representations and use a trainable linear mapping to project it into a single representation, to obtain the model "cat". We observe that they are inferior to our full model, which demonstrates the simplicity and effectiveness of using the addition. A possible reason is that directly adding two representations can effectively maintain their information, and does not introduce any trainable parameters, which may affect the model learning.

Effect of using composed representations as outliers. To further study whether using composed representations as outliers is really beneficial for out-of-distribution detection, we conduct experiments on the UT-Zappos by using real images from the ImageNet dataset [Deng *et al.*, 2009] and the images generated by the stable diffusion [Rombach *et al.*, 2022]. The results are shown Table 3. Specifically, we randomly select $22,998$ images in the ImageNet or use the stable diffusion to generate over $2,2998$ images for the unseen compositions of the UT-Zappos, and require the model to output uniform distribution for these images. The obtained models ("Real images" and "Generated images") underperform our full model. The main reason is that by using composed features, we can obtain features of unseen pairs to better identify unseen compositions. These results demonstrate that adding primitive representations of training images to compose representations of unseen compositions is simple but effective.

4.4 Runtime Consideration

In this part, we analyze the computation time of the grid search. Table 4 shows the time for training one epoch, finding the best threshold T, finding the best fusion weights λ_s, λ_u, and a simplified grid search (SGS) to find the fusion weights, on all datasets. For the SGS, we only perform the grid search within the full range after the first epoch, and then perform the grid search in a small neighborhood around the fusion weights found in the last epoch. For example, if the found λ_u, λ_s are 0.3 and 0.7, the range of the next grid search should be $[0.2, 0.3, 0.4]$ and $[0.6, 0.7, 0.8]$, respectively. It is shown that finding the threshold takes almost no time while finding the fusion weights indeed takes some time. For the MIT-States and C-GQA, finding the fusion weights with the SGS can save a lot of time. Besides, using the SGS leads to the same results with using vanilla grid search, because the fusion weights found of different epochs remain stable. Thus, the extra computations of the grid search are acceptable.

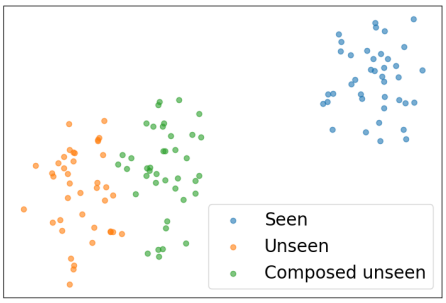

Figure 3: Feature distributions on CGQA.

4.5 Qualitative Results

We visualize the feature distributions of our method with CLIP on the CGQA dataset in Fig. 3 to demonstrate the effectiveness of composing unseen compositions as outliers. Specifically, we select 10 most frequent seen/unseen compositions on the train/test set, respectively, randomly select 4 images for each composition, and visualize their image representations via the t-SNE tool [Van der Maaten and Hinton, 2008]. We also visualize 40 composed representations containing the 4 unseen compositions, by using primitive representations of training images. We observe that composed representations for unseen compositions are close to representations of testing images with unseen compositions, and thus can help our model to identify images with unseen compositions in testing. Note that we only expect composed representations to be beneficial for OOD detection, rather than matching the real distribution of unseen compositions.

5 Conclusion and Future Work

In this work, we have presented COMO, a novel compositional zero-shot learning method. COMO composes representations of unseen compositions as outliers to identify unseen compositions. By combining composition/primitive classification modules, the method is able to separately recognize images with seen/unseen compositions. Extensive experiments show the effectiveness of our method.

In our implementation, the threshold and fusion weights are found via grid search on a validation set, which introduces extra computations. Besides, the fusion weights are the same for all samples with seen/unseen compositions. In the future, we will use several neural modules to predict sample-specific scalars based on each sample, and use reinforcement learning to enable end-to-end training.

References

[Atzmon *et al.*, 2021] Yuval Atzmon, Felix Kreuk, Uri Shalit, and Gal Chechik. A causal view of compositional zero-shot recognition. *Advances in neural information processing systems (NeurIPS)*, 2021.

[Bao *et al.*, 2023] Wentao Bao, Lichang Chen, Heng Huang, and Yu Kong. Prompting language-informed distribution for compositional zero-shot learning. *arXiv preprint arXiv:2305.14428*, 2023.

[Bao *et al.*, 2024] 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, 2024.

[Chen *et al.*, 2020] Xingyu Chen, Xuguang Lan, Fuchun Sun, and Nanning Zheng. A boundary based out-of-distribution classifier for generalized zero-shot learning. In *European Conference on Computer Vision (ECCV)*, pages 572–588. Springer, 2020.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[Fodor and Pylyshyn, 1988] Jerry A Fodor and Zenon W Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[Hendrycks *et al.*, 2018] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations (ICLR)*, 2018.

[Hu and Wang, 2024] Xiaoming Hu and Zilei Wang. A dynamic learning method towards realistic compositional zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2265–2273, 2024.

[Hu *et al.*, 2021] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2021.

[Hu *et al.*, 2023] Yongli Hu, Lincong Feng, Huajie Jiang, Mengting Liu, and Baocai Yin. Domain-aware prototype network for generalized zero-shot learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 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.

[Isola *et al.*, 2015] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *CVPR*, 2015.

[Jiang and Zhang, 2024] Chenyi Jiang and Haofeng Zhang. Revealing the proximate long-tail distribution in compositional zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2498–2506, 2024.

[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.

[Jung *et al.*, 2021] Sanghun Jung, Jungsoo Lee, Daehoon Gwak, Sungha Choi, and Jaegul Choo. Standardized max logits: A simple yet effective approach for identifying unexpected road obstacles in urban-scene segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15425–15434, 2021.

[Karthik *et al.*, 2022] Shyamgopal Karthik, Massimiliano Mancini, and Zeynep Akata. 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)*, pages 9336–9345, 2022.

[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 (CVPR)*, pages 11316–11325, 2020.

[Li *et al.*, 2022] Xiangyu Li, Xu Yang, Kun Wei, Cheng Deng, and Muli Yang. Siamese contrastive embedding network for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9326–9335, 2022.

[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 (CVPR)*, pages 1782–1791, 2023.

[Liu *et al.*, 2021a] Yang Liu, Lei Zhou, Xiao Bai, Yifei Huang, Lin Gu, Jun Zhou, and Tatsuya Harada. Goal-oriented gaze estimation for zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3794–3803, 2021.

[Liu *et al.*, 2021b] Zhe Liu, Yun Li, Lina Yao, Xianzhi Wang, and Guodong Long. Task aligned generative meta-learning for zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8723–8731, 2021.

[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 (CVPR)*, pages 23560–23569, 2023.

[Mancini *et al.*, 2022] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Learning graph embeddings for open world compositional zero-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[Min *et al.*, 2020] Shaobo Min, Hantao Yao, Hongtao Xie, Chaoqun Wang, Zheng-Jun Zha, and Yongdong Zhang. Domain-aware visual bias eliminating for generalized zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12664–12673, 2020.

[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 (CVPR)*, 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 (CVPR)*, pages 953–962, 2021.

[Nayak *et al.*, 2023] Nihal V. Nayak, Peilin Yu, and Stephen Bach. Learning to compose soft prompts for compositional zero-shot learning. In *International Conference on Learning Representations (ICLR)*, 2023.

[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.

[Pourpanah *et al.*, 2022] Farhad Pourpanah, Moloud Abdar, Yuxuan Luo, Xinlei Zhou, Ran Wang, Chee Peng Lim, Xi-Zhao Wang, and QM Jonathan Wu. A review of generalized zero-shot learning methods. *IEEE transactions on pattern analysis and machine intelligence*, 2022.

[Purushwalkam *et al.*, 2019] Senthil Purushwalkam, Maximilian Nickel, Abhinav Gupta, and Marc'Aurelio Ranzato. Task-driven modular networks for zero-shot compositional learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 3593–3602, 2019.

[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 (ICML)*, pages 8748–8763. PMLR, 2021.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

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

[Van der Maaten and Hinton, 2008] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[Verma *et al.*, 2021] Vinay Kumar Verma, Ashish Mishra, Anubha Pandey, Hema A Murthy, and Piyush Rai. Towards zero-shot learning with fewer seen class examples. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2241–2251, 2021.

[Wang *et al.*, 2023a] Henan Wang, Muli Yang, Kun Wei, and Cheng Deng. Hierarchical prompt learning for compositional zero-shot recognition. In *IJCAI*, volume 1, page 3, 2023.

[Wang *et al.*, 2023b] 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 (CVPR)*, pages 11197–11206, 2023.

[Yang *et al.*, 2022] Muli Yang, Chenghao Xu, Aming Wu, and Cheng Deng. A decomposable causal view of compositional zero-shot learning. *IEEE Transactions on Multimedia*, 2022.

[Yu and Grauman, 2014] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *CVPR*, 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 (ECCV)*, pages 339–355. Springer, 2022.

[Zheng *et al.*, 2023] Zhaoheng Zheng, Haidong Zhu, and Ram Nevatia. Caila: Concept-aware intra-layer adapters for compositional zero-shot learning. *arXiv preprint arXiv:2305.16681*, 2023.