Leveraging Sub-class Discimination for Compositional Zero-Shot Learning

Xiaoming Hu, Zilei Wang*

University of Science and Technology of China, Hefei, China
cjdc@mail.ustc.edu.cn, zlwang@ustc.edu.cn

Abstract

Compositional Zero-Shot Learning (CZSL) aims at identifying unseen compositions composed of previously seen attributes and objects during the test phase. In real images, the visual appearances of attributes and objects (primitive concepts) generally interact with each other. Namely, the visual appearances of an attribute may change when composed with different objects, and vice versa. But previous works overlook this important property. In this paper, we introduce a simple yet effective approach with leveraging sub-class discrimination. Specifically, we define the primitive concepts in different compositions as sub-classes, and then maintain the sub-class discrimination to address the above challenge. More specifically, inspired by the observation that the composed recognition models could account for the differences across sub-classes, we first propose to impose the embedding alignment between the composed and disentangled recognition to incorporate sub-class discrimination at the feature level. Then we develop the prototype modulator networks to adjust the class prototypes *w.r.t.* the composition information, which can enhance sub-class discrimination at the classifier level. We conduct extensive experiments on the challenging benchmark datasets, and the considerable performance improvement over state-of-the-art approaches is achieved, which indicates the effectiveness of our method. Our code is available at https://github.com/hxm97/SCD-CZSL.

Introduction

Current deep learning based image recognition algorithms heavily rely on the enormous amount of manually-labeled data. However, for real-world applications where the visual images follows long-tailed distribution, gathering the supervisions for all classes becomes infeasible. Differently, if a person has seen the images of a 'colorful car' and an 'old building', he can understand the concept of a 'colorful building' even without having seen it previously. Therefore, the researchers are committed to integrating such Compositional Zero-shot Learning (CZSL) capability to the computer vision systems (Nan et al. 2019; Huang et al. 2021).

Prior studies have proposed the generalized CZSL setting (Purushwalkam et al. 2019), where the model is evaluated on both seen and unseen compositions. In this paper, we follow such a setting as it requires the model to simultaneously learn new concepts and preserve the discrimination for seen compositions, thus providing a more comprehensive evaluation. In addition, CZSL has the inductive and transductive settings. The inductive setting only uses the training data while the transductive setting uses the unlabeled test data in the training stage (Xu, Kordjamshidi, and Chai 2021). We only take the inductive setting into consideration since the test data is typically unavailable during training.

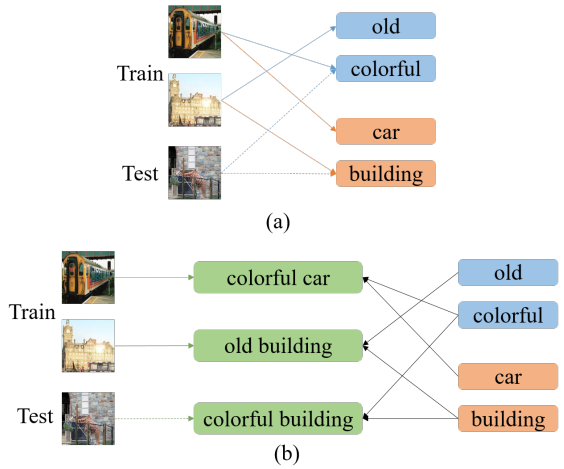

Figure 1: Example of compositional zero-shot classification. (a) Given training samples of colorful car and old building, disentangled CZSL aims at learning an attribute recognition branch (orange line) and an object recognition branch (green line). (b) The composed CZSL learns to generate word embeddings of compositions using NLP techniques (black line), and learns a visual recognition model for compositions (green line). The dotted line represents the recognition branches for unseen images.

The main challenge in CZSL tasks is that the appearances of attributes and objects highly interact with each other in practical scenarios (Yang et al. 2020). As shown in Figure 1, the attribute 'colorful' appears differently in 'colorful car' and 'colorful building', since the car is orange and blue, while the building is grey and brown. Besides, the 'building' represents either as a church in the 'old building' or as

*Corresponding author

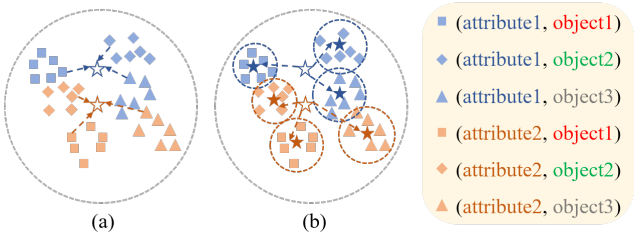

Figure 2: Method motivation: a toy example of attribute classification branch. Different colors indicate samples with different attribute label, while different shapes indicate samples in different sub-class. The pentagrams represent the attribute classification prototypes. (a) Previous methods aim at learning fixed prototype for each class, which would lead to misclassification. (b) Our proposed method pulls the samples within a sub-class together, and learns a dynamic prototype for each sub-class. The object classification branch is tackled similarly. Best viewed in color.

a bounding wall for 'colorful building'. For simplicity, we refer to each attribute and object as a primitive concept, and the term 'sub-class' is defined by considering the same concept present in various compositions. Hence, for the CZSL task, the discrimination among both classes and sub-classes should be taken into account.

We also illustrate two mainstream CZSL approaches of previous works in Figure 1, which we name the disentangled CZSL recognition and composed CZSL recognition, respectively. The disentangled CZSL recognition methods firstly embed the visual features into the attribute domain and object domain separately, and their predictions are then combined to generate the composition prediction in the test stage. Differently, the composed CZSL recognition typically leverages the word embedding of attributes and objects, *e.g.*, glove word vectors (Pennington, Socher, and Manning 2014), to construct the classifier for the training compositions. A similar strategy could be used to create the classifier for unseen compositions.

Although these approaches have achieved promising performance on the CZSL datasets (Nayak, Yu, and Bach 2022; Ruis, Burghouts, and Bucur 2021; Saini, Pham, and Shrivastava 2022), the visual variances between sub-classes have not yet been addressed. On the one hand, the disentangled CZSL recognition methods utilize the concept classifiers to draw the features towards their classification prototype. However, as images of different sub-classes have diverse appearances, pulling their features together would hamper the model generalization for distinguishing unseen sub-classes. On the other hand, the composed CZSL recognition methods partially address this problem in the visual domain by using a composition classifier to distinguish visual features across sub-classes. Nevertheless, the word embedding in the semantic domain is held constant for different input images. For example, the attribute concept 'colorful' shares the same embedding between a 'colorful car' and a 'colorful building'. As a result, such sub-class discrimination cannot be ensured in the semantic domain.

In this work, we propose a simple yet effective method to tackle the interaction between attributes and objects. As shown in Figure 2, we firstly encourage the features within different sub-classes to separate rather than cluster together. In this way, we hope to improve feature-level sub-class discrimination. Then we make the classifier for identifying 'colorful' be distinct for the input images of 'colorful car' and 'colorful building'. That is, we learn to adjust the classification prototypes *w.r.t.* the sub-class representation of primitive concepts. By combining the aforementioned two techniques, we could leverage the sub-class discrimination at both feature-level and classifier-level.

Specifically, at the feature level, we firstly generate virtual attribute and object encodings using two decoders, and then pull the disentangled embedding towards the virtual encoding using contrastive learning. Furthermore, we build an auxiliary composition encoder to produce virtual composite encoding with the disentangled embedding. By forcing the composed encoding to be correctly classified, the appearance variances among compositions would be maintained in the initial disentangled embedding. At the classifier level, we construct two prototype modulators to instruct the classifier to adjust in accordance to the sub-class information. Finally, by combining the above techniques, we incorporate the strengths of both disentangled and composed CZSL recognition methods. Experimental results demonstrate that our method outperforms the state-of-the-art methods by a significant margin. Also, the ablation study confirms that each proposed module can improve the model performance.

In summary, this work presents the following contributions:

- **Sub-class discrimination at feature level:** As the composed classification aims at capturing subtle variances between sub-classes, we apply three auxiliary losses to enhance the alignment between the disentangled embedding and composite embedding.
- **Sub-class discrimination at classifier level:** After the features are pulled towards their composite embedding, the classifier need be adjusted accordingly. To this end, we develop two modulators to modify the classifier prototypes, which takes advantage of sub-class discrimination at the classifier level.
- **State-of-the-art results on benchmark datasets:** We evaluate our model on the benchmark datasets with leveraging sub-class discrimination at both feature-level and classifier-level. Extensive experiments demonstrate the superiority of our method over state-of-the-art methods, as well as the rationality of our model.

Related Work

Disentangled CZSL Recognition Methods

Several recent works train separate recognition models for attribute and objects, and combine their predictions in the test stage. Among them, VisProd (Karthik, Mancini, and Akata 2021) constructs independent fully-connected classification networks for attributes and objects, and shows that this simple baseline could achieve comparable or superior results *w.r.t.* the SOTA CZSL approaches. To further

enforce complete disentanglement of the two recognition branches, the causality independence is utilized in (Atzmon et al. 2020). Moreover, a message passing mechanism is introduced by BMPNet (Xu et al. 2021) to capture relationships between the primitive concepts of attributes and objects. Apart from the methods mentioned above that symmetrically process attributes and objects, Attop (Nagarajan and Grauman 2018) proposes to model attributes as transformation operators, which could transform the object embedding into the appropriate appearance. In addition, inspired by group axioms, SymNet (Li et al. 2020) enforces the symmetry regularization of the object representations given transformations modeled by the attributes. Recently, OADis (Saini, Pham, and Shrivastava 2022) proposes the affinity module for improved disentanglement, which could identify the most similar features of two images with the same primitive concepts. In addition to developing an STM module to generate virtual samples to enhance model generalization, SCEN (Li et al. 2022) uses contrastive learning to capture prototypes for both attribute and object domains. These approaches, however, regard all attributes and objects as consistent within a class, thus neglecting to account for the visual variances between sub-classes. In this work, we propose to tackle this issue with help of the composed CZSL recognition branch.

Composed CZSL Recognition Methods

Previous works commonly use the word embedding of primitive concepts to build the classifier for attribute-object compositions, with a joint compatibility function conditioned on the image, attribute, and object. The typical word embedding approaches include Glove (Pennington, Socher, and Manning 2014), word2vec (Mikolov et al. 2013), and fasttext (Joulin et al. 2016). LabelEmbed (Misra, Gupta, and Hebert 2017) leverages a transformation network to predict the composition classifier parameters. Following a similar idea, TMN (Purushwalkam et al. 2019) learns a modular network whose output compatibility score is reliant on the input image. Several works also construct GAN (Wei et al. 2019) or VAE (Anwaar, Pan, and Kleinsteuber 2022) models to generate virtual features given the semantic representation of an input sample. More recently, CGE (Naeem et al. 2021) uses Graph Neural Network (GNN) to capture the dependence relationship between attributes and objects.

Despite of the fact that the aforementioned methods have achieved competitive performance, they still have two shortcomings. First, they could preserve the sub-class discrimination in the visual domain, but the discrimination in the semantic domain is lost since semantic representations of primitive concepts are kept constant. Second, the performance of these models highly depends on the initialization of word embedding (Saini, Pham, and Shrivastava 2022), whereas an optimal embedding is not always available. In this work, without acquiring any specific natural language processing techniques, we present a simple yet effective CZSL method, which takes the advantages of composed recognition in sub-class discrimination and disentangled recognition in identifying unseen compositions.

Our Approach

The architecture of our proposed approach is illustrated in Figure 3. It consists of two main ingredients: disentangled CZSL recognition (upper stream) and composed CZSL recognition (bottom stream). The design ethos of our method seek to leverage the sub-class discrimination from the composed recognition to improve the disentangled recognition, where the embedding alignment and prototype modulation are proposed. We jointly optimize the both branches in an end-to-end manner. During inference, only the disentangled CZSL recognition branch is used for prediction. In the following section, we first present the CZSL task formulation, and then elaborate on the the baseline framework, embedding alignment module, and prototype modulation module.

Task Formulation

Compositional Zero-shot Learning aims at recognizing novel compositions without having seen their training samples. In this setting, the label of each input image is composed of two primitive concepts, *i.e.*, an attribute and an object. Given the attribute set A and the object set O, we can produce the composition set $C = A \times O = \{(y_a, y_o) \mid y_a \in A, y_o \in O\}$. Suppose that x^i denotes the i^{th} image sample and y_a^i, y_o^i represents its attribute and object label, while $y_c^i = (y_a^i,\ y_o^i)$ represents its composition label. Note that the composition label set $\{y_c^i\}$ for the seen images C_S and unseen images C_U is not overlapped.

Here we consider the more challenging generalized CZSL setting, where the performance for seen and unseen compositions need be balanced. That is, the test set contains the images from both seen and unseen compositions. Overall, the full dataset is typically split into the training set $D_{tr} = \{x_{tr}^i,\ y_c^i\}$, seen test set $D_{sts} = \{x_{sts}^i,\ y_c^i\}$, and unseen test set $D_{uts} = \{x_{uts}^i,\ y_c^i\}$. Especially, all the y_a in A and y_o in O should be included in D_{tr}.

Baseline Framework

Given an image x^i, its visual feature f^i is extracted by the ResNet18 (He et al. 2016) backbone network,then f^i is sent into three encoders, and we name them E_a for attribute encoder, E_o for object encoder, and E_c for composition encoder. Each encoder converts f^i into the embedding in the corresponding domain, resulting in z_a^i, z_o^i, and z_c^i. Finally, we construct the classifiers for each domain, each of which generates predictions for the corresponding domain:

$$L_{cls}^m = -\log \frac{\exp(z_m^i \cdot \mathbf{p}(y_m^i))}{\sum_{y_m} \exp(z_m^i \cdot \mathbf{p}(y_m))}, \tag{1}$$

where $m \in \{a, o, c\}$, and $\mathbf{p}(y_m^i)$ represents the classification prototype vector for y^i in the m^{th} domain. We combine three standard cross entropy losses to jointly optimize these encoders and classifiers:

$$L_{base} = L_{cls}^a + L_{cls}^o + L_{cls}^c. \tag{2}$$

We refer to the attribute and object classification components as the disentangled recognition branch, and the composition classification component as the composed recognition branch. While only the disentangled branch is used

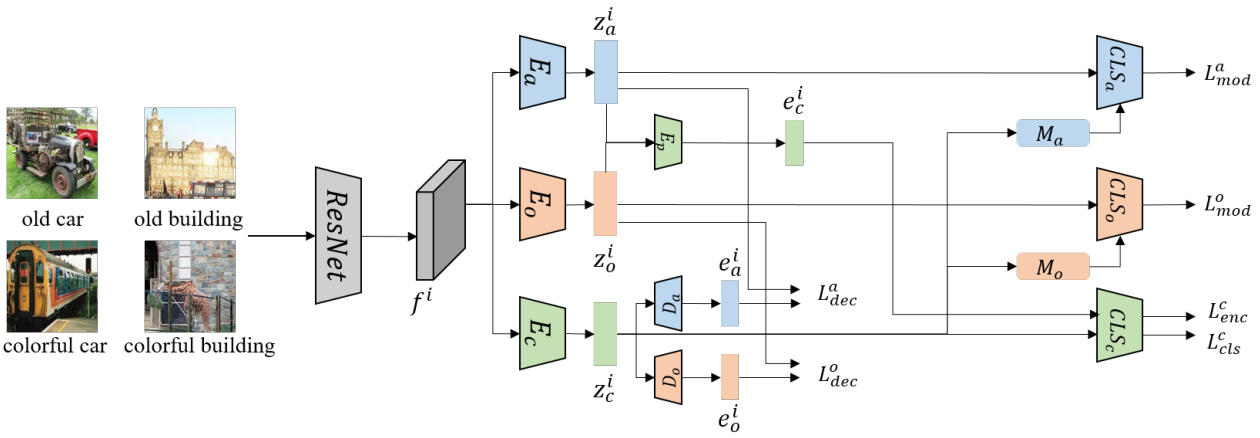

Figure 3: Illustration of our approach. (1) The feature map f^i is first extracted by the backbone network, and then sent to the attribute, object, and composition encoder respectively. (2) The auxiliary encoder E_p generates virtual composite encoding with attribute and object embedding z_a^i and z_o^i as inputs, while the attribute and object decoder D_a and D_o convert the composite embedding z_c^i to the corresponding encoding. (3) The modulator $\mathbf{M}_a$ and $\mathbf{M}_o$ adjust the class prototypes for the attribute and object classification using the composite embedding z_c^i. Best viewed in color.

for inference, our baseline framework is similar to VisProd (Karthik, Mancini, and Akata 2021).

Embedding Alignment at Feature Level

In this section, we present how to integrate sub-class discrimination into the disentangled recognition branch at the feature level. In light of the fact that the composition classification learns to distinguish the concepts in various pairs, we think the embedding of composed recognition could preserve sub-class discrimination. Thus taking the composed embedding z_c^i with two decoders D_a and D_o, we generate the virtual encoding for attribute domain e_a^i and object domain e_o^i. Compared to z_a^i and z_o^i, e_a^i and e_o^i contain more information concerning sub-class discrimination. Here we use contrastive learning (Chen et al. 2020) to impose the embedding alignment between z^i and e^i. To be specific, we set e_a^i as the positive anchor, and randomly select N negative samples from the training batch with different composition labels. We denote the negative sample set as $D_{neg}^i = \{j \mid y_c^j \neq y_c^i\}$. Then we target to decrease the distance between z_a^i to the positive anchor e_a^i, while increasing the distance with negative samples $\{e_a^j\}$, *i.e.*,

$$L_{dec}^a = -\log \frac{\exp((z_a^i \cdot e_a^i)/\tau)}{\exp((z_a^i \cdot e_a^i)/\tau) + \sum_{j \in D_{neg}^i} \exp((z_a^i \cdot e_a^j)/\tau)}, \tag{3}$$

where $\tau > 0$ is a temperature parameter that controls effects of the similar negative samples (Wang and Liu 2021). Note that D_{neg}^i also includes the samples with the same attribute class as x^i but having a different sub-class. For simplicity, we denote these negative samples within the same class as semi-negatives. Impact of such semi-negatives will be discussed in the experiment section.

The alignment loss in object domain L_{dec}^o is processed in the same way with the same negative sample set D_{neg}^i. Then we denote the decoder alignment loss $L_{dec} = L_{dec}^a + L_{dec}^o$. Moreover, we design an auxiliary encoder E_p to generate the composition encoding e_c^i from the concatenation of disentangled embedding z_a^i and z_o^i. In order to further impose the embedding alignment, we hope that e_c^i can also be correctly classified by the composition classifier. We use cross entropy as the encoder alignment loss to integrate sub-class discrimination for z_a^i and z_o^i, *i.e.*,

$$L_{enc} = -\log \frac{\exp(e_c^i \cdot \mathbf{p}(y_c^i))}{\sum_{y_c} \exp(e_c^i \cdot \mathbf{p}(y_c))}. \tag{4}$$

The above techniques are combined in our embedding alignment module, and the integral alignment loss is defined by

$$L_{align} = L_{dec} + L_{enc}. \tag{5}$$

Prototype Modulation at Classifier Level

As the features of different sub-classes constitute various clusters, their classification prototypes should be distinct from each other. However, since each sub-class may only contain a limited number of samples in the training set, learning the classification prototypes directly for each sub-class would degrade model discrimination. Here we develop two modulators $\mathbf{M}_a$ and $\mathbf{M}_o$ for the attribute classification and object classification individually, inspired by (Chou, Lin, and Liu 2020). We take the attribute classification branch as an example in this section. The attribute modulator, $\mathbf{M}_a$, learns to map attribute prototypes towards the corresponding sub-class clustering centers, where $\mathbf{P}_a$ indicates the predefined attribute prototypes, $\mathbf{P}_a = \{\mathbf{p}(y_a) | y_a \in A\}$. Our modulator makes use of the composite embedding z_c^i to adjust the prototypes to represent the sub-class information. Specifically, we refine the classification prototypes by conducting the following operation:

$$\mathbf{P}'_a = \mathbf{P}_a + \mathbf{P}_a \otimes softmax(z_c^i), \tag{6}$$

Dataset	Train	Val		Test	
	sc	sc	uc	sc	uc
UT-Zappos	83	15	15	18	18
C-GQA	5592	1252	1040	888	923

Table 1: Detailed statistics of the used CZSL datasets in our experiments. Here we report the number of seen compositions sc for training split, seen compositions sc and unseen compositions uc for validation and test split from left to right.

where softmax is utilized to produce sub-class attention, $\otimes$ represents Hadamard product to combine the sub-class attention and attribute prototypes. Note that we apply a shortcut operation after the Hadamard product as the softmax function gives large weights on sparse dimensions. Without the shortcut, the important information on sub-class discrimination would be missed.

After prototype modulation, the attribute classification loss L^a_{cls} is modified to

$$L^a_{mod} = -\log \frac{\exp(z^i_a \cdot \mathbf{p}'(y^i_a))}{\sum_{y_a \in A} \exp(z^i_a \cdot \mathbf{p}'(y_a))}. \tag{7}$$

Training and Inference

During training, we use the classification losses with prototype modulation as well as the alignment loss, the overall training loss is denoted by

$$L_{total} = L_{mod} + L^c_{cls} + \alpha \cdot L_{align}, \tag{8}$$

where $L_{mod} = L^a_{mod} + L^o_{mod}$, and α controls the effect of alignment loss. At the test stage, we derive the prediction by choosing the composition which yields the highest prediction score, *i.e.*,

$$\hat{y}^i_c = \arg \max_{(y_a, y_o)} s_a(z^i_a) \cdot s_o(z^i_o). \tag{9}$$

where $s_a(z^i_a) = z^i_a \cdot \mathbf{p}'(y_a)$ and $s_o(z^i_o) = z^i_o \cdot \mathbf{p}'(y_o)$ refer to the classification scores for attribute and object recognition respectively. Here $(y_a, y_o) \in Y_s \cup Y_u$ for the Generalized CZSL setting.

Experiments

Experimental Setting

Dataset Description. We evaluate our method on two benchmark CZSL datasets, *i.e.*, UT-Zappos (Yu and Grauman 2014) and C-GQA (Naeem et al. 2021). Specifically, UT-Zappos is a medium-sized dataset composed of 50025 images of shoes with 16 attribute categories and 12 object categories. Among them, 22998 images are used for training, 3214 for validation, and 2914 for test, respectively. In UT-Zappos, the object label reflects the type of shoes while attribute annotation indicates their material, *e.g.*, faux-leather sandals and wool slippers. Although MIT-States (Isola, Lim, and Adelson 2015) is also a common CZSL dataset, earlier studies (Atzmon et al. 2020) pointed out that due to the infancy of image search engine technology, about 70% samples are mistakenly labeled. As a replacement, we adopt a larger dataset with cleaner label annotation, C-GQA, which is composed of 413 attribute categories and 674 object categories. We use the same data split as proposed in (Purushwalkam et al. 2019) and (Mancini et al. 2022). The detailed statistics of these datasets are summarized in Table 1.

Evaluation Metrics. When testing on both seen and unseen compositions according to the Generalized CZSL setting, there exists significant inductive bias, making the model susceptible to predicting unseen compositions as seen ones. Thus, to balance the model performance over seen and unseen compositions, we adopt the calibrated stacking which lowers the seen class confidence by multiplying a balancing coefficient.

We adopt the same evaluation protocol as prior works, which computes the Area Under the Curve (AUC) of seen-unseen accuracy curve by adjusting the balancing coefficient. The best harmonic mean (HM) between seen and unseen accuracy is also reported. In addition, the best accuracy for seen classes (S) and unseen classes (U) are recorded separately. On UT-Zappos, we conduct our experiment with three different seeds of random number, and report the average precision with error bars to illustrate that the performance of our model is consistent. On CGQA, we only report one accuracy despite the fact that the model's performance on this sizeable dataset is insensitive to random seeds. In particular, AUC and HM are two overall metrics that we mainly concern.

Implementation Details. For fair comparison, we adopt ResNet18 (He et al. 2016) model pretrained on the ImageNet (Russakovsky et al. 2015) as the backbone feature extractor by following previous works. Convolutional layers, batch normalization layers, and fully-connected layers with ReLU activation are used to construct the attribute encoder E_a, object encoder E_o, and composition encoder E_c. Also, we implement the attribute decoder D_a, object decoder D_o, and auxiliary encoder E_p with two fully-connected layers. The number of negative samples used in L_{dec} is set as $N = 10$. Moreover, we conduct our method with PyTorch (Paszke et al. 2019) on a NVIDIA GTX 2080Ti GPU. The model is trained for 50 epochs using the Adam (Kingma and Ba 2014) optimizer with learning rate of $1e^{-4}$ and weight decay of $5e^{-5}$. The temperature parameter τ is set as 0.05, and the weight of alignment loss α is fixed as 1. The model that performs the best on the validation set is used to generate the final results in Table 2.

Comparison with the State-of-The-Art

With the UT-Zappos and CGQA dataset, several relevant and representative works are chosen for comparison: Attop (Nagarajan and Grauman 2018), LabelEmbed+ (Misra, Gupta, and Hebert 2017), TMN (Purushwalkam et al. 2019), SymNet (Li et al. 2020), CGE (Naeem et al. 2021), CompCos (Mancini et al. 2021), Co-CGE (Mancini et al. 2022), OADis (Saini, Pham, and Shrivastava 2022), and SCEN (Li

Model	UT-Zappos				C-GQA			
	AUC	HM	S	U	AUC	HM	S	U
Attop (Nagarajan and Grauman 2018)	25.9	40.8	59.8	54.2	0.7	5.9	17.0	5.6
LE+ (Misra, Gupta, and Hebert 2017)	25.7	41.0	53.0	61.9	0.8	6.1	18.1	5.6
TMN (Purushwalkam et al. 2019)	29.3	45.0	58.7	60.0	1.1	7.5	23.1	6.5
SymNet (Li et al. 2020)	23.9	39.2	53.3	57.9	2.1	11.0	26.8	10.3
CompCos (Mancini et al. 2021)	26.9	41.1	57.7	62.8	2.6	12.4	28.1	11.2
CGE (Naeem et al. 2021)	26.4	41.2	56.8	63.6	2.3	11.4	28.1	10.1
Co-CGE (Mancini et al. 2022)	29.1	44.1	58.2	63.3	2.8	12.7	29.3	11.9
OADis (Saini, Pham, and Shrivastava 2022)	30.0	44.4	59.5	**65.5**	2.9	13.1	**30.5**	12.5
SCEN(Li et al. 2022)	**32.0**	**47.8**	**63.3**	62.5	2.9	12.4	28.9	12.1
Our Method	31.8±0.1	46.3±0.5	62.3±0.8	64.5±0.3	**3.2**	**14.1**	29.9	**14.5**

Table 2: Comparison with state-of-the-art results: we measure the best area under the curve (AUC), best harmonic mean (HM), best seen (S) and unseen accuracy (U) on UT-Zappos and C-GQA dataset. We conduct our method with three different seeds of random number, and report their average precision with error bars on UT-Zappos. Here we reproduce the OADis method on C-GQA using the public open-source code. The best results are marked in bold.

et al. 2022). As shown in Table 2, we have the following observations.

First, on UT-Zappos, our method achieves comparable result with the current SOTA method-SCEN, and outperforms both the disentangled recognition and composed recognition approaches by a significant margin except SCEN. Compared to the OADis, our model improves the overall AUC and harmonic mean metrics by about 2.0%. Compared to SCEN, our model performs about 1.0% lower for the harmonic mean and seen class accuracy, and about the same for the other two evaluation metrics. Moreover, although generation-based models often obtain higher results as pointed by previous ZSL works, our method is a simple disciminative method without using memory bank (D_{ir} in SCEN) and generator (STM in SCEN), achieving comparable results with far less computational cost. We confirm a similar trend on the C-GQA, where our method surpasses both the current SOTA, Co-CGE, and our reproduced OADis. Specifically, we observe an improved performance in the AUC from 2.9% to 3.2%, and the harmonic mean from 13.1% to 14.1%. Finally, it should be noted that random number seeds have a significant impact on how well a model performs on UT-Zappos, so it is important to report the average precision by choosing a variety of random number seeds. We recommend that future works to proceed in this manner.

Ablation Study

In this section, we ablate our model to illustrate the effectiveness of our proposed modules. As for the baseline model, we train combination of disentangled classification branch and composition classification branch. The disentangled classification branch is then used for testing. On top of this baseline, we gradually add the following techniques:

- With L_{dec}, the composite embedding z_c^i is decomposed into the attribute domain e_a^i and the object domain e_o^i. The training process is guided by L_{base}+L_{dec}.
- With L_{enc}, the attribute representation z_a^i and object representation z_o^i are used to generate composite encoding e_c^i. The training process is guided by L_{base}+L_{enc}.
- With **M**, the prototypes for the attribute and object classifiers are modulated by the composite embedding z_c^i. The training process is guided by $L_{mod} + L_{cls}^c$.

Here we consider all the possible combinations in Table 3. We can see that all of the components consistently improve model performance, and combining them together performs the best. As can be concluded, the improved sub-class discrimination at both feature and classifier level is beneficial for learning unseen compositions.

Negative Sampling Strategy

We use the semi-negatives for decoder alignment, where the samples from the same class may be interpreted as negative samples. Mentioning that SCEN (Li et al. 2022) also develops a contrastive learning mechanism for resolving the CZSL task, it should be noted that ours is somewhat counterintuitive since SCEN only treats samples from various classes as negatives. Here we investigate the impact of different negative sampling strategies to validate its rationality. We control the number of two types of negative samples. Taking the attribute alignment module as example, here we sample K semi-negative samples with the same attribute label and different object label, along with $N - K$ fully-negative samples with different attribute and object labels. Note that the K samples should be sampled from the whole dataset, as some classes might correspond to none semi-negative samples in a training batch. However, sampling from the whole dataset is rather time-consuming. Thus in Table 4, we fix the total number of negative samples as 10, and then select K from $\{0, 1, 2\}$. We can draw the conclusion that the model performance is boosted by raising K from 0 to 2, benefited from the promoted sub-class discrimination. Additionally, we test directly sampling negative samples with distinct sub-class labels in a training batch, where the proportion of semi-negative samples is assigned randomly. In our experiments, we use this sampling strategy due to its compelling performance and low computational cost. It is worth noting that we only use UT-Zappos for experiment in this section since the semi-negative classes are

Method			UT-Zappos				C-GQA			
L_{dec}	L_{enc}	**M**	AUC	HM	S	U	AUC	HM	S	U
✗	✗	✗	29.5±0.3	45.0±0.7	61.6±0.8	63.6±0.6	2.9	13.2	29.1	13.7
✓	✗	✗	30.8±1.0	45.6±0.8	62.4±1.2	64.5±0.4	3.1	13.4	29.4	13.9
✗	✓	✗	30.7±0.5	45.3±0.2	61.9±0.6	64.0±0.5	3.1	13.7	29.1	14.0
✗	✗	✓	30.5±0.7	45.5±0.4	62.3±0.6	**65.0±0.1**	3.0	13.2	29.5	13.9
✓	✓	✗	31.1±0.2	45.6±0.5	62.1±0.7	64.5±0.4	3.1	13.8	29.0	14.1
✓	✗	✓	31.2±0.6	45.5±0.8	62.5±0.5	64.9±0.9	**3.2**	13.6	29.5	14.0
✗	✓	✓	31.0±0.2	45.6±0.5	**63.1±0.7**	64.5±0.4	**3.2**	14.0	29.0	14.2
✓	✓	✓	**31.8±0.1**	**46.3±0.5**	62.3±0.8	64.5±0.3	**3.2**	**14.1**	**29.9**	**14.5**

Table 3: Ablation studies: we quantitatively verify the effectiveness of the proposed losses and modulator by ablating over the architecture of our model. Best results are marked in bold.

Method		UT-Zappos			
N	K	AUC	HM	S	U
10	0	31.2±0.5	45.7±0.8	**62.3±1.0**	64.6±0.4
10	1	31.4±0.3	45.8±0.3	62.2±1.0	**65.0±1.0**
10	2	31.6±0.2	45.8±0.0	62.0±1.0	64.8±0.5
10	∼	**31.8±0.1**	**46.3±0.5**	**62.3±0.8**	64.5±0.3

Table 4: Negative Sampling Strategies: we quantitatively verify rationality of using semi-negative samples in feature alignment module. Best results are marked in bold.

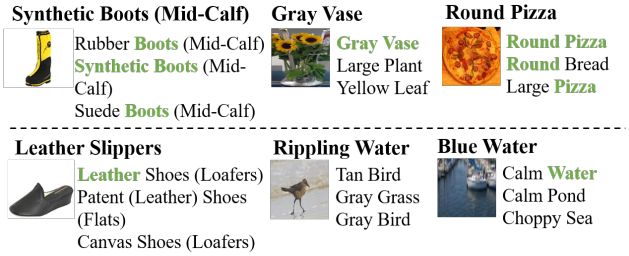

Figure 4: Qualitative results of top-3 attribute and object predictions for test samples. The ground truth label of each test sample is shown on top of each image, while the top-3 attribute and object predictions are listed on the right side. The concepts with correct predictions are marked in green.

not available for some training classes on C-GQA, as earlier works (Saini, Pham, and Shrivastava 2022) pointed out.

Qualitative Results

We show several qualitative results for novel compositions with their predictions. For the first row in Figure 4, we present the images whose top-3 predictions match the ground-truth label on UT-Zappos and C-GQA respectively. It can be seen that our model consistently provides accurate composition predictions, which confirms the superiority of leveraging the advantage of sub-class discrimination. As for the failure cases in the last row, this is primarily caused by the issue of incomplete annotation. That is, multiple attributes and objects may exist in an image although only one attribute and object is annotated. Then it is rather difficult for our method to choose which attribute and object to forecast on. For instance, the 'calm' attribute also appears in the 'blue water' image, while the 'rippling water' image includes a gray bird in the center. To the best of our knowledge, it still remains an unsolved problem to deal with such incomplete annotation challenge for the CZSL task.

Limitations

Despite of the superior performance, our method still has some flaws. First, although our method could reduce the computational cost of composed recognition models due to removing the dependence on the word embedding of primitive concepts. However, this limits its open-set application when the probable test data compositions are unknown in advance. This problem could be addressed by incorporating word embedding into the disentangled recognition model training process (Karthik, Mancini, and Akata 2022). Second, the same concept in various sub-classes sometimes maintain a consistent visual representation, for example, a blue car and a green car might have the same visual appearance for 'car'. Nonetheless, our method increases the distance between their representations under such circumstances. In future works, the sub-class discrimination remains to be explored more precisely.

Conclusion

In this paper, we develop a simple yet effective method to recognize unseen attribute-object compositions. First, we employ the composite embedding as positive anchors, imposing the disentangled recognition branch to maintain sub-class discrimination at the feature level. Then we design two modulators to dynamically modify the classifier prototypes towards the cluster center of sub-class features. We verified the effectiveness of our proposed method on benchmark datasets, where comparison experiments illustrate that it outperforms previous state-of-the-art approaches, and ablation studies validate the rationality of our model. Finally, we also discuss the limitations of our work, which we hope will motivate future works to better overcome the CZSL challenges.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant 62176246 and Grant 61836008.

References

Anwaar, M. U.; Pan, Z.; and Kleinsteuber, M. 2022. On Leveraging Variational Graph Embeddings for Open World Compositional Zero-Shot Learning. *arXiv preprint arXiv:2204.11848*.

Atzmon, Y.; Kreuk, F.; Shalit, U.; and Chechik, G. 2020. A causal view of compositional zero-shot recognition. *Advances in Neural Information Processing Systems*, 33: 1462–1473.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chou, Y.-Y.; Lin, H.-T.; and Liu, T.-L. 2020. Adaptive and generative zero-shot learning. In *International conference on learning representations*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Huang, H.; Tang, W.; Zhang, J.; and Yu, P. S. 2021. Translational Concept Embedding for Generalized Compositional Zero-shot Learning. *arXiv preprint arXiv:2112.10871*.

Isola, P.; Lim, J. J.; and Adelson, E. H. 2015. Discovering states and transformations in image collections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1383–1391.

Joulin, A.; Grave, E.; Bojanowski, P.; and Mikolov, T. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Karthik, S.; Mancini, M.; and Akata, Z. 2021. Revisiting visual product for compositional zero-shot learning. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*.

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*, 9336–9345.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Li, X.; Yang, X.; Wei, K.; Deng, C.; and Yang, M. 2022. Siamese Contrastive Embedding Network for Compositional Zero-Shot Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9326–9335.

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*, 11316–11325.

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*, 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*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

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*, 1792–1801.

Naeem, M. F.; Xian, Y.; Tombari, F.; and Akata, Z. 2021. Learning graph embeddings for compositional zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 953–962.

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

Nan, Z.; Liu, Y.; Zheng, N.; and Zhu, S.-C. 2019. Recognizing unseen attribute-object pair with generative model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 8811–8818.

Nayak, N. V.; Yu, P.; and Bach, S. H. 2022. Learning to Compose Soft Prompts for Compositional Zero-Shot Learning. *arXiv preprint arXiv:2204.03574*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

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.

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.

Ruis, F.; Burghouts, G.; and Bucur, D. 2021. Independent prototype propagation for zero-shot compositionality. *Advances in Neural Information Processing Systems*, 34: 10641–10653.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252.

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*, 13658–13667.

Wang, F.; and Liu, H. 2021. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2495–2504.

Wei, K.; Yang, M.; Wang, H.; Deng, C.; and Liu, X. 2019. Adversarial fine-grained composition learning for unseen attribute-object recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3741–3749.

Xu, G.; Kordjamshidi, P.; and Chai, J. Y. 2021. Zero-shot compositional concept learning. *arXiv preprint arXiv:2107.05176*.

Xu, Z.; Wang, G.; Wong, Y.; and Kankanhalli, M. S. 2021. Relation-aware Compositional Zero-shot Learning for Attribute-Object Pair Recognition. *IEEE Transactions on Multimedia*.

Yang, M.; Deng, C.; Yan, J.; Liu, X.; and Tao, D. 2020. Learning unseen concepts via hierarchical decomposition and composition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10248–10256.

Yu, A.; and Grauman, K. 2014. Fine-grained visual comparisons with local learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 192–199.