A Dynamic Learning Method towards Realistic 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

To tackle the challenge of recognizing images of unseen attribute-object compositions, Compositional Zero-Shot Learning (CZSL) methods have been previously addressed. However, test images in realistic scenarios may also incorporate other forms of unknown factors, such as novel semantic concepts or novel image styles. As previous CZSL works have overlooked this critical issue, in this research, we first propose the Realistic Compositional Zero-Shot Learning (RCZSL) task which considers the various types of unknown factors in an unified experimental setting. To achieve this, we firstly conduct re-labelling on MIT-States and use the pre-trained generative models to obtain images of various domains. Then the entire dataset is split into a training set and a test set, with the latter containing images of unseen concepts, unseen compositions, unseen domains as well as their combinations. Following this, we show that the visual-semantic relationship changes on unseen images, leading us to construct two dynamic modulators to adapt the visual features and composition prototypes in accordance with the input image. We believe that such a dynamic learning method could effectively alleviate the domain shift problem caused by various types of unknown factors. We conduct extensive experiments on benchmark datasets for both the conventional CZSL setting and the proposed RCZSL setting. The effectiveness of our method has been proven by empirical results, which significantly outperformed both our baseline method and state-of-the-art approaches.

Introduction

Current computer vision algorithms generally learn to recognize the object category of an given image. However, a single object may be coupled with multiple attributes, making it difficult for computer vision systems to generalize their object recognition capabilities to unseen images. For instance, training dataset might include images of "peeled apple" and "ripe apple", while we may require to identify "sliced apple" during test. Since gathering supervisions for all available attributes is infeasible, researchers are committed to studying Compositional Zero-shot Learning (CZSL) task which aims at recognizing novel compositions of seen attribute and object concepts. Furthermore, previous studies have proposed the transductive CZSL setting (Xu, Kordjamshidi, and Chai 2021) and partial supervision CZSL setting (Karthik, Mancini, and Akata 2022). These settings consider a special circumstance of CZSL, when the unlabelled test images could be obtained or partial labels are absent during training. Actually, we assert that during training, labels are typically intact and test data is unavailable, thus restricting the practical applications of these settings.

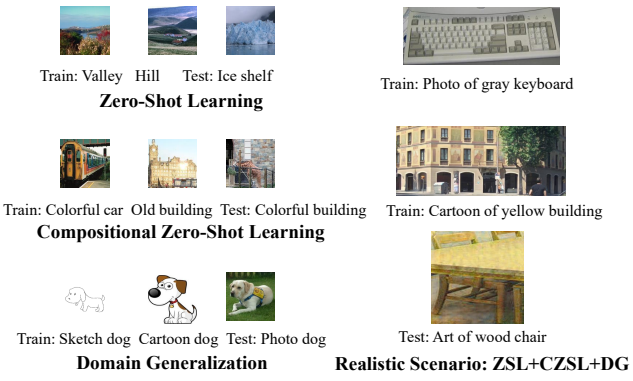

Figure 1: Zero-Shot Learning (ZSL), Compositional Zero-Shot Learning (CZSL) and Domain Generalization (DG) tasks require to recognize unseen semantic categories, unseen compositional categories and unseen image styles respectively. In contrast, the goal of Realistic Compositional Zero-Shot Learning (RCZSL) task is to identify images that concurrently contain all the aforementioned types of unknown factors.

While prior CZSL researches have made great progress, we propose that a crucial issue has been overlooked. In previous CZSL settings, the unseen compositions are considered to be composed of seen attributes and seen objects. We demonstrate that this is not often the case in real-world situations. The requirement for human annotations restricts the capacity of training datasets, therefore test images may also contain previously unseen attribute and object concepts or unseen styles. To address this, Zero-Shot Learning (ZSL) and Domain Generalization (DG) approaches have been pro-

*Corresponding author

posed. Nevertheless, such issue has not yet been raised in the CZSL tasks.

We illustrate the distinction between ZSL, CZSL and DG tasks in Figure 1. As can be observed, ZSL and CZSL tasks aim at recognizing unknown object concepts or unknown attribute-object compositions where the test images follow the same style pattern as training images. While in DG tasks, the test images have the same object concept as the training images but with a distinct style. Consequently, there are semantic concepts, attribute-object compositions, and the image styles that make up the complete list of unknown factors.

To this end, we first propose the Realistic Compositional Zero-Shot Learning (RCZSL) setting in which the test images contain all three types of unknown factors as well as their combinations. Additionally, we adopt the generalized setting following prior ZSL and CZSL works as both seen and unseen categories could be present during the test phase (Chao et al. 2016; Purushwalkam et al. 2019). Therefore the test set can be separated into six distinct groups based on the type of unknown factors contained. Specifically, there are test images with seen styles and unseen styles, both including seen compositions, unseen compositions with seen concepts, and unseen compositions with unseen concepts, respectively. The three types of unknown factors are simultaneously present in a single test image which are comprised of unseen attributes, unseen objects and unseen styles, as an extreme circumstance in Figure 1. We demonstrate that our proposed RCZSL setting provides a more comprehensive evaluation, and is more applicable in realistic scenarios.

The main challenge in RCZSL tasks lies in that there lacks available datasets and evaluation metrics. To address this, we first conduct re-labelling for MIT-States (Isola, Lim, and Adelson 2015) as former researches have pointed out that most images were mistakenly labelled (Atzmon et al. 2020). We fix the incorrect label annotations to create a more reasonable benchmark. Besides, we leverage the pretrained Generative Adversarial Network (GAN) models to transform images in MIT-States into various domains. Moreover, we divide the the entire dataset of images into the training, validation and test set, where the latter two sets are further divided into the aforementioned six distinct groups. We also construct the evaluation metric by computing the harmonic mean of the recognition accuracy in various test image groups.

We contend that in the RCZSL setting, the three types of unknown factors would lead to the domain shift problem. In particular, a certain attribute or object concept might have different visual representations, while the visual appearances of attributes and objects also depend on each other. Moreover, the image styles could also induce the visual appearance variation. Such a domain shift problem would induce bias to the visual-semantic relationship on test images, thus decreasing the recognition accuracy.

In this work, we propose a dynamic learning method to address the domain shift issue. At the visual feature level, we firstly combine the style and semantic components of an image to obtain the visual embedding, via stacking the low-level convolutional feature statistics, and conducting Global Average Pooling (GAP) on the high-level feature respectively. The coefficient prediction module is subsequently created to produce the weight of fixed convolution kernels. Finally the dynamic convolution module adaptively generates convolution kernels according to the predicted coefficients, thus adjusting the model to novel domains and compositions.

Moreover, we offer two optional approaches to generate the semantic prototypes of compositions, either by leveraging the text encoder of CLIP model (Yun et al. 2022) or using the Object Conditioned Network (OCN) (Saini, Pham, and Shrivastava 2022) to combine the Glove (Pennington, Socher, and Manning 2014) word vector of attributes and objects. Then the low-level visual features that have been scaled to the same size are stacked before inputting into the semantic modulator. The original composition prototypes are updated dynamically by localizing related regions *w.r.t.* the input image. Combining the aforementioned techniques, our method may bridge the domain shift brought on by unknown factors by modulating the visual features and composition prototypes dynamically. Experimental results show that on the conventional CZSL setting and the proposed RCZSL setting, our method significantly outperforms the baseline and state-of-the-art methods.

In summary, this work presents the following contributions:

- **A Realistic Compositional Zero-Shot Learning Setting:** We propose a more realistic setting that integrates the unseen concept categories and unseen image styles into the conventional CZSL setting. By developing the benchmark dataset and evaluation metrics, we hope this setting could encourage future works to perform better in practical applications.
- **A Dynamic Learning Method for RCZSL:** We further develop a dynamic learning method that incorporates the visual and semantic modulators. The domain shift issue could thus be alleviated by altering the visual features and the composition prototypes in accordance with the input image.
- **State-of-the-art results on benchmark datasets:** We evaluate our model on both CZSL and RCZSL tasks. Extensive experiments back up the effectiveness of our dynamic learning method by illustrating that it outperforms both baseline and state-of-the-art methods.

Related Work

Compositional Zero-Shot Learning

Previous CZSL methods can be divided into the disentangled recognition methods and the composed recognition methods. The former approaches, on the one hand, train distinct recognition models for attributes and objects, and then integrate their predictions during the test stage. Among them, VisProd (Karthik, Mancini, and Akata 2021) builds independent, fully-connected classification networks for attributes and objects, and indicates that this simple baseline could produce results that are on par with or even better than those produced by SOTA CZSL techniques. Attop (Nagarajan and Grauman 2018) proposes modeling attributes as

transformation operators, which could change the appearance of the object embedding. Moreover, SymNet (Li et al. 2020) enforces the symmetry regularization of the object representations given transformations described by the attributes, drawing inspiration from group axioms. SCEN (Li et al. 2022) employs a STM module to generate virtual samples as well as a contrastive learning mechanism to capture attribute and object prototypes. Recently, to simulate the interactions between different primitives, DRANet (Li et al. 2023b) employs the reverse-and-distill approach that disentangles the attribute and object embeddings.

On the other hand, the composed recognition approaches 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. A transformation network is used by LabelEmbed (Misra, Gupta, and Hebert 2017) to predict the composition classifier's input parameters. An analogous approach is used by TMN (Purushwalkam et al. 2019), which develops a modular network whose output compatibility score is dependent on the input image. To generate virtual features given the semantic representation of an input sample, several works also construct the GAN (Wei et al. 2019) or Variational Auto-Encoder (VAE) (Anwaar, Pan, and Kleinsteuber 2022) models. Also, CGE (Naeem et al. 2021) makes use of Graph Neural Network (GNN) to depict the interdependence of attributes and objects. More recently, CANet (Wang et al. 2023b) utilizes a hyper learner and a base learner to learn attributes conditioned on the recognized object. Note that there are also methods that combine disentangled recognition and composed recognition in an unified framework (Hu and Wang 2023).

Zero-Shot Learning

ZSL methods can be divided into embedding-based or generation-based approaches, where the embedding-based ZSL works mainly consider to use different embedding space (Zhang, Xiang, and Gong 2017), embedding function (Akata et al. 2015), and regularization term (Kodirov, Xiang, and Gong 2017) to perform knowledge transfer. Recently, the potential of local representations in embedding-based ZSL methods has been investigated. AREN (Xie et al. 2019) uses the second-order pooling to create the discriminative features and the attention layer to localize various regions. In order to improve the discrimination of image representations, APN (Xu et al. 2020) employs a prototype learning branch to promote image features to incorporate more local information. PSVMA (Liu et al. 2023) utilizes the dual semantic-visual transformer module to achieve semantic-visual mutual adaptation for semantic disambiguation. On the other hand, the generation-based methods are to convert ZSL into a fully supervised classification problem by generating samples of target classes. GAN models (Xian et al. 2018; Su et al. 2022), VAE models (Schonfeld et al. 2019; Wang et al. 2023c), and generative flows (Shen et al. 2020) have shown good capability to generate virtual image features for unseen classes.

Domain Generalization

Aiming at recognizing unseen domains, prior domain generalization methods construct domain-invariant features via adversarial learning (Kim, Li, and Hospedales 2023) or meta learning (Qin, Song, and Jiang 2023) mechanisms. In order to further enhance the generalization capability of the model, the recent methods also introduce a causality regularizer (Chen et al. 2023), gradient matching network (Wang et al. 2023a), or trajectory sampling (Wang, Grigsby, and Qi 2023) into the baseline meta-learning framework. Furthermore, researchers adopt data augmentation methods to create samples similar to target domains. For example, GAN models are used by L2A-OT (Zhou et al. 2020) to create new images that are distributed differently from the original domain but semantically consistent. CycleMAE (Yang et al. 2023) uses a Masked Auto-Encoder (MAE) model to create a cycle image reconstruction task that results in more realistic and unique image pairs. Other typical domain generalization methods generally include model ensemble (Zhou et al. 2021; Qu et al. 2022), feature normalization (Zhu et al. 2022; Meng et al. 2022) or contrastive learning techniques (Yao et al. 2022; Li et al. 2023a).

Our Approach

In this work, we either use the CLIP-free or the CLIP-based method, where the former makes use of ResNet18 and Glove as the visual feature and semantic prototype extractors, while the latter utilizes pre-trained ViT-L/14 as both visual and semantic backbone. Note that the CLIP-based method only includes the semantic modulator since it is unable to conduct dynamic convolution on the 1-d visual features generated by ViT backbone networks. We illustrate the architecture of the CLIP-free method in Figure 2 as an example. In the following section, we first present the CZSL and RCZSL task formulation, and then elaborate on the the baseline framework, visual modulator, and the semantic modulator module respectively. Note that our method follows the same training and inference paradigm as the baseline method under both CZSL and RCZSL setting.

Task Formulation

In this work, we are given the attribute set A, the object set O and the domain set D. Suppose that x^i denotes the i^{th} image sample and y_a^i, y_o^i, y_d^i represents its attribute, object and domain label, while $y^i = (y_a^i, y_o^i, y_d^i)$ represents its intact label. Thus the label annotation set is decided by $Y = A \times O \times D = \{(y_a, y_o, y_d) \mid y_a \in A, y_o \in O, y_d \in D\}$. Overall, the full dataset is typically split into the training set D_{tr}, the validation set D_{vl} and the test set D_{te}.

We consider both the CZSL as well as the RCZSL settings. For the CZSL setting, all images have the same y_d^i, while all the y_a^i and y_o^i from the validation and test sets are also present in the training set. That is to say, all the y_a and y_o should be included in D_{tr}. For the RCZSL setting, the y_a^i, y_o^i and y_d^i labels on the validation and test sets could be either seen or unseen during training. In this setting, the test images are referred to as unseen compositions with unseen concepts when they contain unseen y_a^i or unseen y_o^i, whereas

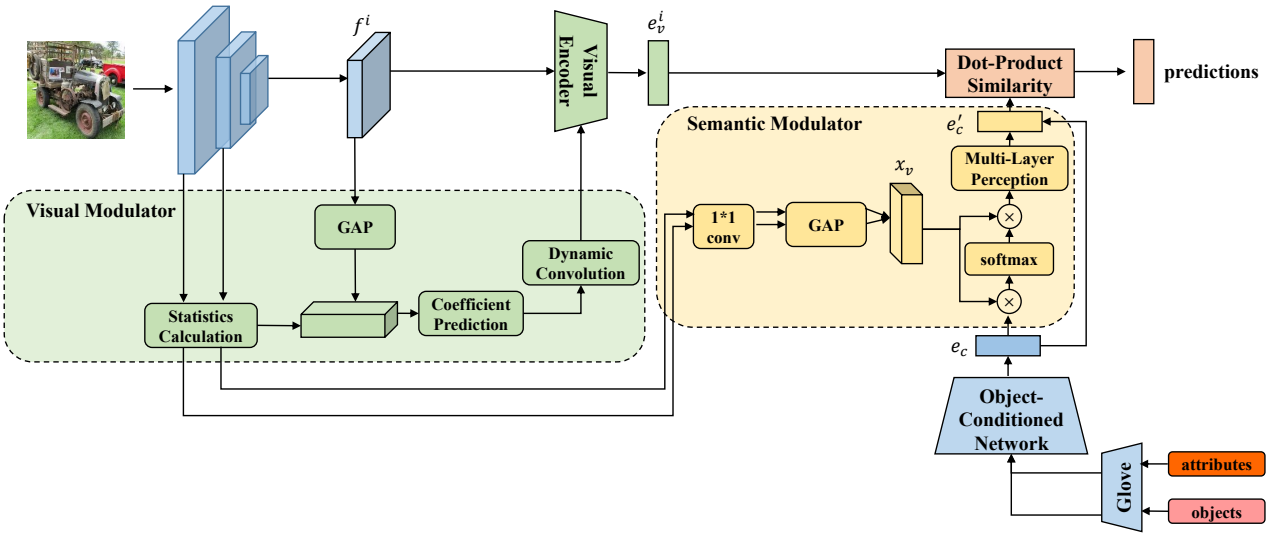

Figure 2: Illustration of our approach. (1) The feature map f^i is first extracted by the backbone network, and then sent to visual encoder. (2) The OCN module combines the word vectors of attributes and objects, obtaining the semantic prototypes for compositions. (3) The two modulators adjust the visual features and semantic prototypes respectively. Best viewed in color.

unseen compositions with seen concepts occur when both y_a^i and y_o^i are seen in the training phase, but are combined in a different manner.

Baseline Framework

Given the input 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 the visual encoder to obtain its visual embedding e_v^i. Such visual encoder is composed of convolution layer, Batch Normalization (BN) and the GAP. In addition, the Glove word vectors of attributes v_a and objects v_o are combined by the OCN module to generate the composition prototypes e_c for both seen and unseen compositions. Here we derive the predictions by calculating the dot product similarity between e_v^i and e_c, and choose the available composition which yields the highest prediction score:

$$\hat{y}_c^i = \arg\max_{(y_a,y_o)} e_v^i \cdot e_c, \tag{1}$$

where $\cdot$ represents the dot product operation. $(y_a, y_o) \in D_{tr}$ during training while $(y_a, y_o) \in D_{tr} \cup D_{vl} \cup D_{te}$ during test. Finally we use the cross entropy loss to optimize the visual and semantic encoders:

$$L_{cls} = -\log \frac{\exp(e_v^i \cdot e_c^i)}{\sum_{e_c \in E_c} \exp(e_v^i \cdot e_c)}, \tag{2}$$

where e_c^i represents the composition prototype composed of y_a^i and y_o^i, $\mathbf{E_c} = \{e_1, ..., e_{|C|}\}$ represents the prototype set of all compositions with $|C|$ representing the number of possible compositions.

Dynamic Visual Modulator

In this section, we present how to dynamically adjust the visual encoder thus the domain shift problem could be alleviated at the visual feature level. In light of the fact that the unknown factors can be split into semantic-related (unseen concepts and unseen compositions) and style-related (unseen domains) ones, we leverage the semantic and style embedding to perform dynamic convolution. Specifically the semantic embedding x_{sem}^i is obtained by conducting GAP directly on f^i, whereas the style embedding x_{sty}^i is achieved via concatenating on the feature statistics produced by the blocks in the backbone ResNet. For instance, given the feature x_j^i generated by the j-th block, the style statistics are calculated across spatial dimensions for each channel:

$$\mu(x_j^i) = \frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} x_j^i \tag{3}$$

$$\sigma(x_j^i) = \sqrt{\frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} (x_j^i - \mu(x_j^i))^2 + \epsilon}, \tag{4}$$

where H and W representing height and width of x_j^i. Afterwards, the semantic embedding x_{sem}^i is obtained by stacking $\mu(x_j)$ and $\sigma(x_j)$ for $j \in \{1, 2, 3, 4\}$. Then the semantic embedding and style embedding are concatenated and sent into the coefficient prediction module which is designed as a Multi-Layer Perception (MLP). The coefficient prediction module provides weights for the convolution kernels, then the multiple kernels are aggregated by the weights to dynamically adjust the visual features f^i. Such dynamic convolution is used to replace the static convolution of the visual encoder in the baseline method.

Dynamic Semantic Modulator

In this section, we show how the semantic modulator dynamically adjusts the composition prototypes generated by

Figure 3: Re-Labelling of MIT-States. We show the original label annotations present in MIT-States as well as our correction, with green indicating the correct label, red indicating that the original label disagrees with our correction.

Dataset	Train	Val		Test	
	sc	sc	uc	sc	uc
MIT-States	1262	300	300	400	400
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.

the OCN for the CLIP-free method or the text encoder for the CLIP-based method. In order to capture the visual variances of the same composition in various images, the features generated by each block of the backbone ResNet are stacked after resize to make up the visual representation x_v of an image. To this end, these features are resized to the same shape using 1*1 convolution and GAP. We choose x_v instead of x_i as the image representation for these low-level features better preserve the visual details, they update E_c by the semantic modulator via:

$$\mathbf{E}'_{\mathbf{c}} = f(x_v * softmax(x_v^T * \mathbf{E}_{\mathbf{c}})) + \mathbf{E}_{\mathbf{c}}, \qquad (5)$$

where is $f(\cdot)$ a MLP network and $*$ represents the matrix multiplication. Finally, the prototypes after modulation could be aware of the visual variances of the same semantic in different images, as they effectively aggregate the visual details in x_v that related to the semantic information of each composition.

Experiments

Experimental Setting

Dataset for CZSL setting. We evaluate our method on three benchmark CZSL datasets, *i.e.*, MIT-States (Isola, Lim, and Adelson 2015), UT-Zappos (Yu and Grauman 2014) and C-GQA (Naeem et al. 2021). Specifically, MIT-States is a common CZSL dataset composed of 53753 images with 115 attribute categories and 245 object categories. In MIT-States, 30338 images are used for training, 10420 images for validation, and 12995 for test. 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. We also adopt C-GQA which is composed of 413 attribute categories and 674 object categories. In C-GQA, the number of images used for training, validation and test are 26920, 7280 and 5098 respectively. We use the same data split as proposed in (Purushwalkam et al. 2019) and (Mancini et al. 2022). The detailed statistics of the used CZSL datasets are summarized in Table 1.

Num	Train	Val	Test
	se	$<se, uc, up>$	$<se, uc, up>$
pair	1398	-, 108, 74	-, 193, 189
img	22548	4330, 3240, 1978	10103, 5604, 5950
domain	2	3, 3, 3	3, 3, 3

Table 2: Detailed statistics of the created MIT-States-RCZSL dataset in our experiments. From left to right: the seen compositions on the training split (se), the seen compositions (se), unseen compositions with seen concepts (uc) and with unseen concepts (up) on the validation split and on the test split.

Dataset for RCZSL setting. Here we propose that there is no existing dataset for the proposed RCZSL task. Therefore we developed the RCZSL benchmark dataset based on the MIT-States, in which the natural images included could better reflect the real-world circumstances. We note that earlier studies (Atzmon et al. 2020) have pointed out that due to the infancy of image search engine technology, about 70% samples in MIT-States are mistakenly labeled. To tackle this, we first perform re-labelling on MIT-States to create a cleaner and more believable experimental setting. Figure 3 illustrates that many images in MIT-States are initially allocated with the incorrect label, and how our re-labelling more accurately describes the image. We encourage future works to follow this benchmark, for the incorrect label annotations could lead to the opposite conclusion.

We then apply the pre-trained style transfer model to acquire images from other domains, namely art and cartoon, as MIT-States only consists images of the photo domain. To be more precise, we apply the CycleGAN model (Zhu et al. 2017) to conduct photo-to-art transfer, which is pre-trained using adversarial and cycle-consistency losses. Additionally, we perform photo-to-cartoon transfer using an another GAN model (Wang and Yu 2020) that individually recognizes surface, structure, and texture representations of cartoons.

Then the generated MIT-States-RCZSL dataset has 53753 images and 1962 pairs for each of the three domains which are split into training, validation and test sets. The training set specifically consists of two training domains, each with 1398 pairs and 22420 images. The validation set and test

Model	MIT-States				UT-Zappos				C-GQA			
	AUC	HM	S	U	AUC	HM	S	U	AUC	HM	S	U
Attop (Nagarajan and Grauman 2018)	1.6	9.9	14.3	17.4	25.9	40.8	59.8	54.2	0.7	5.9	17.0	5.6
LE+ (Misra, Gupta, and Hebert 2017)	2.0	10.7	15.0	20.1	25.7	41.0	53.0	61.9	0.8	6.1	18.1	5.6
TMN (Purushwalkam et al. 2019)	2.9	13.0	20.2	20.1	29.3	45.0	58.7	60.0	1.1	7.5	23.1	6.5
SymNet (Li et al. 2020)	3.0	16.1	24.4	25.2	23.9	39.2	53.3	57.9	2.1	11.0	26.8	10.3
CompCos (Mancini et al. 2021)	4.5	16.4	25.3	24.6	26.9	41.1	57.7	62.8	2.6	12.4	28.1	11.2
CGE (Naeem et al. 2021)	5.1	17.2	28.7	25.3	26.4	41.2	56.8	63.6	2.3	11.4	28.1	10.1
Co-CGE (Mancini et al. 2022)	5.1	17.5	27.8	25.2	29.1	44.1	58.2	63.3	2.8	12.7	29.3	11.9
DeCa (Yang et al. 2022)	5.3	18.2	29.8	25.2	31.6	46.3	62.7	63.1	-	-	-	-
SCEN (Li et al. 2022)	5.3	18.4	29.9	25.2	32.0	47.8	63.3	62.5	2.9	12.4	28.9	12.1
SCD (Hu and Wang 2023)	4.8	17.6	30.7	24.6	31.8	46.3	62.3	64.5	3.2	14.1	29.9	**14.5**
CANet (Wang et al. 2023b)	5.4	17.9	29.0	26.2	33.1	47.3	61.0	66.3	**3.3**	14.5	30.0	**14.5**
Our Method	**5.8**	**19.2**	**30.7**	**26.6**	**37.7**	**52.1**	**66.5**	**68.1**	**3.3**	**14.8**	**30.7**	**14.5**
CLIP (Yun et al. 2022)	11.0	26.1	30.2	46.0	5.0	15.6	15.8	49.1	1.4	8.6	7.5	25.0
CSP (Nayak, Yu, and Bach 2022)	19.4	36.3	**46.6**	**49.9**	33.0	46.6	64.2	66.2	6.2	20.5	28.8	26.8
Our Method	**20.0**	**37.4**	46.3	49.8	**39.6**	**52.0**	**67.1**	**72.5**	**7.3**	**21.9**	**32.4**	**28.5**

Table 3: 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 MIT-States, UT-Zappos and C-GQA dataset. Here the results produced by CLIP-free and CLIP-based methods are separately reported for a fair comparision. The best results are marked in bold.

set both comprise three groups of images, which we refer to them as the seen pairs during training, the unseen compositions composed by seen concepts, and the unseen pairs composed by unseen concepts during training, represented by se, uc and up, respectively. Detailed description of the division of MIT-States-RCZSL can be found in Table 2.

Evaluation Metrics. For CZSL setting, there exists significant inductive bias when testing on both seen and unseen compositions, 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.

While for the RCZSL setting, the image recognition accuracy on the three groups for each domain should all be considered. Here we introduce two balancing coefficients which adds on the prediction scores for unseen compositions of seen concepts and unseen concepts respectively. Thus the highest harmonic mean of the three groups' accuracy would be achieved via adjusting the balancing coefficients. In the subsequent experiment, each group's maximum accuracy and their maximum harmonic mean are given separately, while harmonic mean is the main concern.

Implementation Details. We conduct our method with the PyTorch (Paszke et al. 2019) framework. Here we provide the results of the CLIP-free method by adopting the pre-trained ResNet18 model to generate the visual features and Glove algorithm to generate the composition prototypes. Also we conduct the CLIP-based models by leveraging pre-trained ViT-L/14 models as both visual encoder and text encoders. 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 number of dynamic kernels in the visual modulator is set as 4. The visual encoder and text encoder in CLIP and the ResNet18 backbone are fixed during the experiment for fair compasision with prior works. The model that performs the best on the validation set is used to produce the final test results.

Experimental Results for CZSL

With the MIT-States, UT-Zappos and CGQA dataset, several representative works are chosen for comparison. As shown in Table 3, we conclude that on the three datasets, our method outperform the current SOTA methods-SCEN and CANet by a significant margin. We claim that such gain can be attributed to that our method addressed the domain shift introduced by the attribute-object interaction. Especially, our method significantly surpasses state-of-the-art methods on UT-Zappos, improving the harmonic mean and overall AUC metrics by nearly 5.0%. Moreover, we also achieve comparable results on the MIT-States and C-GQA.

In comparision with the CLIP-based methods, we offer the results produced by the fixed CLIP model (Yun et al. 2022) as well as the CSP method (Nayak, Yu, and Bach 2022) which uses the soft prompt mechanism to generate the composition prototypes. It can be concluded that our method also outperforms these approaches with the proposed semantic modulator.

Experimental Results for RCZSL

In this section, we use the CLIP-free method to conduct the experiment without loss of generality. The art and cartoon domains are separately chosen as the test domain, for the photo domain has been observed at the model pre-training stage, thus would violate the "unseen image style" principle if chosen as the test domain. Here we ablate our model architecture to illustrate the effectiveness of the visual and se-

Test Domain	Modulator		Train Domain 1				Train Domain 2				Test Domain			
	Vis	Sem	*se*	*uc*	*up*	HM	*se*	*uc*	*up*	HM	*se*	*uc*	*up*	HM
Art	✗	✗	27.0	43.0	33.6	19.7	26.3	42.5	31.4	18.9	16.8	33.2	25.1	12.8
	✓	✗	41.0	49.3	37.0	26.8	40.0	48.5	35.8	26.2	23.8	36.5	26.4	16.0
	✗	✓	36.0	47.8	36.7	24.6	34.6	46.4	35.4	23.0	22.7	36.9	28.5	16.1
	✓	✓	41.3	49.4	38.3	27.5	40.7	48.7	37.1	27.1	24.9	36.2	27.3	16.7
Cartoon	✗	✗	29.5	44.7	33.9	21.0	26.3	40.6	31.9	18.8	24.4	37.5	28.4	17.0
	✓	✗	39.3	48.7	36.5	26.4	37.2	46.5	35.3	25.2	29.9	41.8	31.1	19.8
	✗	✓	35.0	47.1	34.9	23.6	33.1	45.9	34.9	22.8	27.5	38.8	28.4	18.1
	✓	✓	40.0	49.6	36.4	26.5	37.8	47.6	35.1	25.2	31.0	43.0	31.3	20.7

Table 4: Ablation studies on the RCZSL setting. We quantitatively verify the effectiveness of the visual and semantic modulators by ablating over the architecture of our model. Here *se*, *uc* and *up* respectively represents the recognition accuracy for seen compositions, unseen compositions with seen concepts and unseen concepts, while HM represents their maximum harmonic mean. The train domain 1 stands for the photo domain, while the train domain 2 stands for the rest.

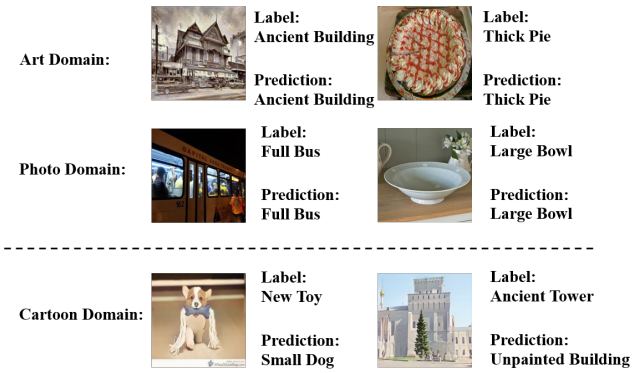

Figure 4: Qualitative results of attribute and object predictions for test samples on MIT-States-RZSL. The ground truth label and the predictions are listed for each image.

mantic modulator. As observed from Table 4, we conclude that both modulators improve the model performance, and combining them together performs the best. Additionally, the accuracy improvement on all the three groups prove that our method effectively tackles the domain shift brought by various types of unknown factors. Finally, we believe that our method builds a baseline for RCZSL thus could motivate more future researches into this setting.

Qualitative Results

We show several qualitative results on MIT-States-RCZSL in Figure 4, where the test images with the predictions of our model are provided. The first column shows the test images with unseen compositions of seen concepts, while the images of the second column are with unseen concepts. In the first two rows, our predictions could accurately describe the visual content of the test images, even though multiple types of unknown factors simultaneously exist in the test images. Also there are failure cases as shown in the third row. We state that such failure cases are partially attributed to the issue of incomplete annotation. The multi-label characteristic of natural images provides additional challenge for the RCZSL task.

Limitations

We provide the RCZSL task that incorporates the novel semantic categories, novel compositions, novel image styles, as well as their combinations which could be present in the real-world unseen images in the meanwhile. Our research, however, only concentrates on the task of image classification, leaving other significant computer vision applications to be investigated. For instance, a Compositional Zero-Shot Segmentation system might play an important role in segmenting novel target structures for medical image analysis, while a Compositional Zero-Shot Video Object Segmentation system could better explore the challenging autonomous driving tasks. In conclusion, we hope that our work could inspire future researches into other real-world Compositional Zero-Shot Learning computer vision settings.

Conclusion

In this paper, we propose a Realistic Compositional Zero-Shot Learning (RCZSL) setting which considers the unseen concepts and unseen styles in the conventional CZSL tasks. We build the corresponding dataset by conducting relabelling and employing pre-trained GAN models to generate images of various domains. The dataset split and evaluation metrics are determined by considering the generalized circumstance. To prevent the domain shift under this setting, we design two dynamic modulators that adapt the visual features and composition prototypes according to the input images respectively. We verified the effectiveness of our proposed method on both the standard CZSL setting and the proposed RCZSL setting. Comparison experiments illustrate that our method outperforms previous state-of-the-art approaches, and ablation studies support the effectiveness of the dynamic learning approach. Finally, we discuss the limitations of our work, which we hope could motivate future researches to explore more practical applications of CZSL.

Acknowledgements

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

References

Akata, Z.; Reed, S.; Walter, D.; Lee, H.; and Schiele, B. 2015. Evaluation of Output Embeddings for Fine-grained Image Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2927–2936.

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.

Chao, W.-L.; Changpinyo, S.; Gong, B.; and Sha, F. 2016. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, 52–68. Springer, Cham: Springer International Publishing. ISBN 978-3-319-46475-6.

Chen, J.; Gao, Z.; Wu, X.; and Luo, J. 2023. Meta-causal Learning for Single Domain Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7683–7692.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Hu, X.; and Wang, Z. 2023. Leveraging Sub-class Discimination for Compositional Zero-Shot Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 890–898.

Isola, P.; Lim, J. J.; and Adelson, E. H. 2015. Discovering states and transformations in image collections. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1383–1391.

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.

Kim, M.; Li, D.; and Hospedales, T. 2023. Domain Generalisation via Domain Adaptation: An Adversarial Fourier Amplitude Approach. *arXiv preprint arXiv:2302.12047*.

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

Kodirov, E.; Xiang, T.; and Gong, S. 2017. Semantic autoencoder for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3174–3183.

Li, C.; Zhang, D.; Huang, W.; and Zhang, J. 2023a. Cross Contrastive Feature Perturbation for Domain Generalization. *arXiv preprint arXiv:2307.12502*.

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.; Liu, Z.; Jha, S.; Cripps, S.; and Yao, L. 2023b. Distilled Reverse Attention Network for Open-world Compositional Zero-Shot Learning. *arXiv preprint arXiv:2303.00404*.

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.

Liu, M.; Li, F.; Zhang, C.; Wei, Y.; Bai, H.; and Zhao, Y. 2023. Progressive Semantic-Visual Mutual Adaption for Generalized Zero-Shot Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15337–15346.

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*, 1–1.

Meng, R.; Li, X.; Chen, W.; Yang, S.; Song, J.; Wang, X.; Zhang, L.; Song, M.; Xie, D.; and Pu, S. 2022. Attention diversification for domain generalization. In *Computer Vision – ECCV 2022*, 322–340. Springer, Cham: Springer Nature Switzerland. ISBN 978-3-031-19830-4.

Misra, I.; Gupta, A.; and Hebert, M. 2017. From red wine to red tomato: Composition with context. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1160–1169.

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 *Computer Vision – ECCV 2018*, 169–185. Cham: Springer International Publishing. ISBN 978-3-030-01246-5.

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 *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 3592–3601.

Qin, X.; Song, X.; and Jiang, S. 2023. Bi-level Meta-learning for Few-shot Domain Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15900–15910.

Qu, J.; Faney, T.; Wang, Z.; Gallinari, P.; Yousef, S.; and de Hemptinne, J.-C. 2022. Hmoe: Hypernetwork-based mixture of experts for domain generalization. *arXiv preprint arXiv:2211.08253*.

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.

Schonfeld, E.; Ebrahimi, S.; Sinha, S.; Darrell, T.; and Akata, Z. 2019. Generalized zero-and few-shot learning via aligned variational autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8247–8255.

Shen, Y.; Qin, J.; Huang, L.; Liu, L.; Zhu, F.; and Shao, L. 2020. Invertible zero-shot recognition flows. In *Computer Vision – ECCV 2020*, 614–631. Springer, Cham: Springer International Publishing. ISBN 978-3-030-58517-4.

Su, H.; Li, J.; Chen, Z.; Zhu, L.; and Lu, K. 2022. Distinguishing unseen from seen for generalized zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7885–7894.

Wang, P.; Zhang, Z.; Lei, Z.; and Zhang, L. 2023a. Sharpness-aware gradient matching for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3769–3778.

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

Wang, X.; and Yu, J. 2020. Learning to cartoonize using white-box cartoon representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8087–8096.

Wang, Z.; Grigsby, J.; and Qi, Y. 2023. PGrad: Learning Principal Gradients For Domain Generalization. *arXiv preprint arXiv:2305.01134*.

Wang, Z.; Hao, Y.; Mu, T.; Li, O.; Wang, S.; and He, X. 2023c. Bi-directional Distribution Alignment for Transductive Zero-Shot Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19893–19902.

Wei, K.; Yang, M.; Wang, H.; Deng, C.; and Liu, X. 2019. Adversarial fine-grained composition learning for unseen attribute-object recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 3740–3748.

Xian, Y.; Lorenz, T.; Schiele, B.; and Akata, Z. 2018. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5542–5551.

Xie, G.-S.; Liu, L.; Jin, X.; Zhu, F.; Zhang, Z.; Qin, J.; Yao, Y.; and Shao, L. 2019. Attentive region embedding network for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9376–9385.

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

Xu, W.; Xian, Y.; Wang, J.; Schiele, B.; and Akata, Z. 2020. Attribute prototype network for zero-shot learning. In *Advances in Neural Information Processing Systems*, volume 33, 21969–21980.

Yang, H.; Li, X.; Tang, S.; Zhu, F.; Wang, Y.; Chen, M.; Bai, L.; Zhao, R.; and Ouyang, W. 2023. Cycle-consistent Masked AutoEncoder for Unsupervised Domain Generalization. In *The Eleventh International Conference on Learning Representations*.

Yang, M.; Xu, C.; Wu, A.; and Deng, C. 2022. A decomposable causal view of compositional zero-shot learning. *IEEE Transactions on Multimedia*, 1–11.

Yao, X.; Bai, Y.; Zhang, X.; Zhang, Y.; Sun, Q.; Chen, R.; Li, R.; and Yu, B. 2022. Pcl: Proxy-based contrastive learning for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7097–7107.

Yu, A.; and Grauman, K. 2014. Fine-grained visual comparisons with local learning. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 192–199.

Yun, T.; Bhalla, U.; Pavlick, E.; and Sun, C. 2022. Do Vision-Language Pretrained Models Learn Primitive Concepts? *arXiv preprint arXiv:2203.17271*.

Zhang, L.; Xiang, T.; and Gong, S. 2017. Learning a deep embedding model for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021–2030.

Zhou, K.; Yang, Y.; Hospedales, T.; and Xiang, T. 2020. Learning to generate novel domains for domain generalization. In *Computer Vision–ECCV 2020*, 561–578. Springer, Cham: Springer International Publishing. ISBN 978-3-030-58517-4.

Zhou, K.; Yang, Y.; Qiao, Y.; and Xiang, T. 2021. Domain adaptive ensemble learning. *IEEE Transactions on Image Processing*, 30: 8008–8018.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.

Zhu, W.; Lu, L.; Xiao, J.; Han, M.; Luo, J.; and Harrison, A. P. 2022. Localized adversarial domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7108–7118.