Continual Compositional Zero-Shot Learning

Yang Zhang[1] , **Songhe Feng**[2,*] , **Jiazheng Yuan**[3,*]
[1]School of Computer Science and Technology, Beijing Jiaotong University, Beijing, China
[2]Tangshan Research Institute, Beijing Jiaotong University, Beijing, China
[3]College of Science and Technology, Beijing Open University, Beijing, China
{23111124, shfeng}@bjtu.edu.cn, jzyuan@139.com

Abstract

Compositional Zero-Shot Learning (CZSL) aims to recognize unseen compositions with the knowledge learned from seen compositions, where each composition is composed of two primitives (attribute and object). However, existing CZSL methods are designed to learn compositions from fixed primitive set, which cannot handle the continually expanding primitive set in real-world applications. In this paper, we propose a new CZSL setting, named Continual Compositional Zero-Shot Learning (CCZSL), which requires the model to recognize unseen compositions composed of learned primitive set while continually increasing the size of learned primitive set. Contextuality and catastrophic forgetting are the main issues to be addressed in this setting. Specifically, we capture similar contextuality in compositions through several learnable Super-Primitives that can modify the invariant primitive embedding to better adapt the contextuality in the corresponding composition. Then we introduce a dual knowledge distillation loss which aims at maintaining old knowledge learned from previous sessions and avoiding overfitting of new session. We design the CCZSL evaluation protocol and conduct extensive experiments on widely used benchmarks, demonstrating the superiority of our method compared to the state-of-the-art CZSL methods.

1 Introduction

Human beings can decompose observations into primitive concepts and recombine these primitive concepts to generalize to unseen compositions. Compositional Zero-Shot Learning (CZSL) [Misra *et al.*, 2017] is proposed to mimic the way humans recognize unseen compositions which is often regarded as a hallmark of intelligence [Atzmon *et al.*, 2016], [Lake *et al.*, 2017], [Li *et al.*, 2023]. Generally, CZSL methods are trained on seen compositions and deployed to recognize unseen compositions, both of which are composed of two primitives (attribute and object) in the fixed set. Such methods are infeasible for many real-world applications as they cannot continually learn and accumulate knowledge of new primitives that might emerge after training. On the contrary, humans can not only recognize unseen compositions, but also learn new concepts incrementally throughout their whole lives.

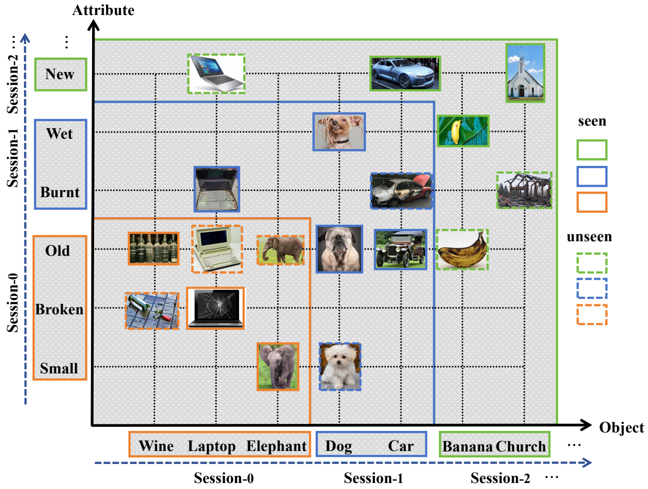

Figure 1: An illustration of our CCZSL setting. We require the model to recognize unseen compositions across multiple incremental sessions, where each session introduces at least one new primitive.

In recent years, sporadic researches have emerged in Continual Generalized Zero-Shot Learning (CGZSL) that require model to continually recognize new unseen categories. [Wei *et al.*, 2020] considers each GZSL dataset as an incremental session and accumulate different attribute knowledge from multiple datasets. However, this method needs task-id during testing phase and only test on individual GZSL dataset, which is difficult to meet in realistic scenarios. Then, [Wu *et al.*, 2023b], [Skorokhodov and Elhoseiny, 2020] and [Gautam *et al.*, 2020] divide a dataset into multiple subsets in different ways to simulate the continual learning process. Although these methods propose various reasonable ways to divide the dataset, compared to the first method, these methods only focus on continual learning of categories and ignore continual learning of attributes. That is to say, the above incremental

[*] Corresponding Authors

sessions may not bring new attribute knowledge to the model. As the categories all come from the same attribute set and there are no explicit constraints on attribute increments for each incremental session.

Drawing on the experience of CGZSL mentioned above, we propose a new Continual Compositional Zero-Shot Learning (CCZSL) setting. This setting requires the model to recognize unseen compositions composed of learned primitive set while continually increasing the size of this set. As illustrated in Figure 1, we set up multiple incremental sessions to bring new primitives step by step. Concretely, the model learns new primitives through new seen compositions in each incremental session and is ready to recognize unseen compositions composed of primitives in the expanded set, that is, unseen compositions of all sessions up to the current session. From another perspective, we split the standard CZSL into multiple CZSL sub-tasks, as newly added primitives and old primitives are combined into several seen compositions and unseen composition like standard CZSL. After learning the last incremental session, the model learns the same number of attributes, objects and compositions in standard CZSL dataset. Next, we will introduce two challenges in this setting, 1) Contextuality and 2) Catastrophic Forgetting, and their corresponding solutions.

Specifically, contextuality indicates that the appearance of the same attribute (object) undergoes significant changes when combined with different objects (attributes). For example, the same attribute “`old`” depicts retro style for object “`car`” but aged appearance for object “`dog`”. The same object “`church`” exhibits a complete structure in composition “`new church`”, yet only damaged structure and debris are bespoke in “`burnt church`”. [Zhang *et al.*, 2022] ignores the contextuality and aims to learn generic primitive embeddings, which are then used to predict the corresponding primitives in all compositions. [Hu and Wang, 2023] learns multiple embeddings for each primitive according to compositions, but is hard to achieve due to data scarcity. Unlike the methods mentioned above, we propose to additionally learn multiple super-primitives which can modify the invariant primitive embedding to better fit contextuality in the current composition. We found that although contextuality significantly varies appearances of primitives, there are still consensuses in these variations. For example, the super-object “`mammal`” beyond “`dog`”, “`cat`” and “`elephant`” has the similar contextuality, like dull fur and sagging skin, when combine with attribute “`old`”. In the same way, the super-attribute “`fragmented`” beyond “`broken`” and “`sliced`” tends to change appearance of object by breaking object into multiple small pieces. Therefore, we propose to learn super-primitives from original primitives, which encapsulate primitives with similar contextuality in compositions. During the inference phase, we recognize attribute (object) by introducing super-object (super-attribute), which can be utilized to simulate contextuality in the current composition and modify the invariant attribute (object) embedding.

Only using samples of new categories to train model will cause forgetting of old knowledge, resulting in a decrease in overall performance, which is known as catastrophic forgetting [Kirkpatrick *et al.*, 2017]. In this setting, as the learned primitives emerge in new compositions with fresh contextuality during subsequent incremental sessions, the model experiences catastrophic forgetting when attempting to relearn these primitives in the new compositions. To be specific, introducing new primitives in incremental sessions leads to the creation of new compositions that incorporate both the new and old primitives. The newfound contextuality within these compositions changes visual appearances of the old primitives, which are distinct from those observed in previous sessions. Learning from the above new compositions will lead to the forgetting of the old primitive characteristics. Data regularization is a popular and concise solution in continual learning, which aims to minimize the impact of new incremental sessions on weights important for previous sessions. In this paper, we propose to utilize knowledge distillation to keep the learned embedding of previous primitives from drifting too much. To maintain the characteristics of networks learned from previous sessions, we adjust the scope of knowledge distillation, introducing constraints on predicting new categories as well.

The main contributions of our work are summarized as follows:

- To the best of our knowledge, we are the first to propose and tackle CCZSL. This setting holds great significance and practicality for real-world applications. Additionally, we design the evaluation protocols and conduct comprehensive experiments.
- We propose to address the issue of contextuality by learning multiple super-primitives corresponding to similar contextuality. Then, we propose a dual knowledge distillation loss to keep old primitive knowledge from catastrophic forgetting and mitigate overfitting on new primitives.
- Experiments and ablation studies affirm the effectiveness of our proposed method.

2 Related Work

2.1 Compositional Zero-Shot Learning

Zero-Shot Learning (ZSL) [Palatucci *et al.*, 2009][Zhang and Feng, 2023][Tang *et al.*, 2020] is a popular research topic that provides model with shared semantic knowledge (attribute knowledge) between seen and unseen classes, enabling them to recognize unseen classes without any training data. CZSL is a sub-topic of ZSL, which requires model to recognize unseen compositions with the shared primitive knowledge. In the early attempts at CZSL, methods try to directly learn the classifiers of compositions by transforming the paired primitive embeddings into composition embedding. For example, [Misra *et al.*, 2017] projects the image representations and paired primitives into a common space, and then predicts unseen compositions in the common space. [Nagarajan and Grauman, 2018] and [Li *et al.*, 2020] treat attributes as operators which change original objects into compositions. [Naeem *et al.*, 2021] tries to exploit the dependency between attributes, objects and their compositions within a graph structure to transfer the relevant knowledge from seen to unseen. Recent mainstream works disentangle visual representations of primitives and learn corresponding embeddings respectively. [Li *et al.*, 2022] leverages contrastive loss to excavate discriminative embeddings of primitives. [Zhang *et al.*, 2022] considers CZSL as a Out-Of-Distribution problem and learns invariant primitive embeddings. [Saini *et al.*, 2022] and [Hao *et al.*, 2023] implement cross-attention between two samples sharing the same primitive to disentangle the corresponding representation of this primitive. [Wang *et al.*, 2023] proposes to learn attribute embeddings that rely on objects and images. [Kim *et al.*, 2023] disentangles attribute representation from feature map with the guidance of object visual representation.

2.2 Continual Generalized Zero-Shot Learning

Continual Generalized Zero-Shot Learning (CGZSL) is the most related work to ours. [Wei *et al.*, 2020] is the first work to investigate GZSL in continual learning scenario. They construct a sequence of tasks by assembling existing GZSL benchmarks and the performance is reported for each benchmark individually. However, this setting requires a task-level supervision in the form of task-ids at test time which is hard to achieve. Subsequently, some follow works propose new CGZSL settings, such as [Skorokhodov and Elhoseiny, 2020] and [Gautam *et al.*, 2020]. The former randomly divides individual dataset into *T* tasks and assumes all previously encountered tasks as seen classes and future tasks as unseen classes. The latter equips an exclusive set of seen and unseen classes for each task and the model can accommodate any number of tasks. Due to the popularity of generative methods [Xian *et al.*, 2018][Tang *et al.*, 2022] in GZSL, generative replay is the mainstream technique for mitigating catastrophic forgetting in the above CGZSL. Although the recent CGZSL settings are more practical, the construction of incremental sessions still has flaws. As attribute plays a important role in GZSL, dividing the dataset based on attribute knowledge is better than random splitting. That is, introducing new attributes for each session, and the new categories are naturally selected from the dataset based on the new attributes.

2.3 Knowledge Distillation

Knowledge distillation (KD) [Hinton *et al.*, 2015] was originally designed to learn a more compact student network from a larger teacher network. Then, it became a major technique for data regularization methods in continual learning. [Li and Hoiem, 2017] first leverages this technique by constructing a cross-entropy loss, known as learning without forgetting loss, to keep representations of previous data from drifting too much while learning new task. This loss was initially employed in task-incremental learning (task-IL) methods [Titsias *et al.*, 2019]. Subsequently, owning to numerous studies observing its effectiveness in settings with small domain shifts between tasks, such as class-incremental learning (class-IL)[Chi *et al.*, 2022], [Wu *et al.*, 2023a], the loss has become a crucial component in many class-IL methods, such as [Dhar *et al.*, 2019] and [Zhang *et al.*, 2020]. It is worthy noting that, as reported in [Masana *et al.*, 2022], continual learning and class-IL are mostly the same.

3 Methodology

3.1 Problem Formulation

CCZSL requires the model to recognize unseen compositions composed of learned primitive set while continually increasing the size of learned primitive set. To simulate the continual learning scenario, we split the original CZSL dataset into *T+1* sessions $\{\mathcal{D}^0, \mathcal{D}^1, ..., \mathcal{D}^T\}$, where each session consists of a training set $\mathcal{D}^i_{tr}$ and a testing set $\mathcal{D}^i_{ts}$. Specifically, let $\mathcal{D}^i_{tr} = \{(x, a, o, c)|x \in \mathcal{X}^i_{tr}, a \in \mathcal{A}^i_{tr}, o \in \mathcal{O}^i_{tr}, c \in \mathcal{C}^i_{tr}\}$, x denotes the image in the sample space of i-th session $\mathcal{X}^i_{tr}$ and c is the composition label in the label space of i-th session $\mathcal{C}^i_{tr}$, each label c is composed of the attribute a and the object o in the tuple respectively, $i.e.$, $c = (a, o)$. In the same way, the testing set is defined as follows $\mathcal{D}^i_{ts} = \{(x, a, o, c)|x \in \mathcal{X}^i_{ts}, a \in \mathcal{A}^i_{ts}, o \in \mathcal{O}^i_{ts}, c \in \mathcal{C}^i_{ts}\}$. From the perspective of increasing sessions, the label spaces of the training set in different sessions are disjoint, $i.e.$, $\mathcal{C}^i_{tr} \cap \mathcal{C}^j_{tr} = \emptyset$, but the label space of the testing set in current session are the summary of label spaces from all previous sessions, $i.e.$, $\forall i < j$, $\mathcal{C}^i_{ts} \subset \mathcal{C}^j_{ts}$. The primitive set follows the same pattern mentioned above in both training and testing set, $i.e.$, $\forall i < j$, $\mathcal{P}^i_{tr/ts} \subset \mathcal{P}^j_{tr/ts}$, $\mathcal{P} = \{\mathcal{A}, \mathcal{O}\}$.

3.2 Preliminary

We first outline the fundamental pipeline of the recent mainstream methods. When provided with an image x, these methods initially utilize a backbone network, $i.e.$, ResNet18, to extract its high-level visual features denoted as X. Subsequently, three feature encoders are respectively employed to disentangle the corresponding attribute, object, and composition representations, which are defined as follows:

$$v_a = E_a(X), v_o = E_o(X), v_c = E_c(X). \tag{1}$$

To train a model capable of understanding primitives and composing them to recognize unseen compositions, a common approach is to directly maximize the similarity scores, such as cosine similarity, between these extracted representations and corresponding embeddings. The function $s(\cdot)$ calculates the similarity scores as follows:

$$s(v_p, p) = \frac{v_p^T}{||v_p||} \cdot \frac{e_p}{||e_p||}, \tag{2}$$

$$s(v_c, (a, o))) = \frac{v_c^T}{||v_c||} \cdot \frac{\varphi_c(e_a, e_o)}{||\varphi_c(e_a, e_o)||}, \tag{3}$$

where $p = \{a, o\}$. v_p and e_p are the corresponding representation and embedding. v_c refers to composition representation. $\varphi_c(\cdot)$ is the transformation network that transforms paired primitive embeddings into composition embedding.

3.3 Capturing Contextuality by Super-Primitives

Generally, relying on contextuality-free and invariant primitive embeddings to recognize the diverse primitive representations in different compositions is evidently insufficient. Therefore, we propose to learn K super-primitives to capture

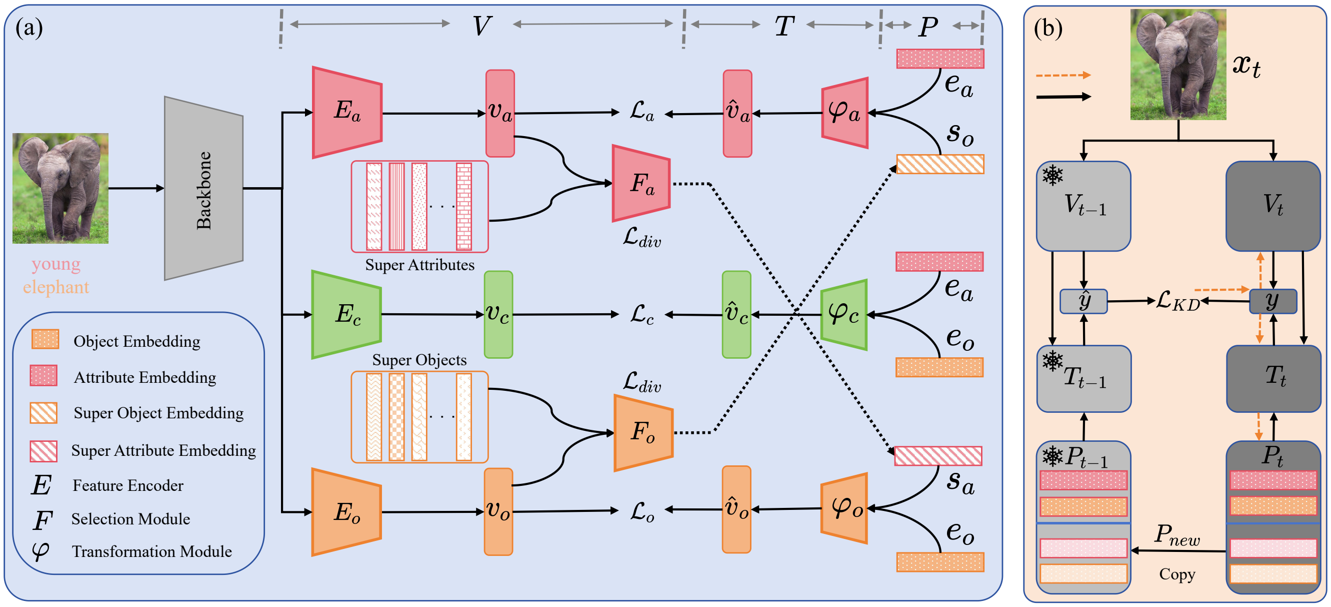

Figure 2: Method overview. (a) The main pipeline of our CCZSL model. Our model can be divided into three components, namely V, T and P. V refers to the primitive representations extraction module, T indicates the transformation module and P includes all the primitive embeddings. Specifically, in V, there are three encoders E_a, E_o and E_c which are used to extract attribute, object and composition representation, *i.e.*, v_a, v_o and v_c. F_a and F_o are super-primitive selection module which select appropriate super-primitive, *i.e.*, s_a and s_o, based on the extracted primitive representation. In T, we have three transformation networks φ_a, φ_o and φ_c, which incorporate contextuality into primitive embeddings, e_a and e_o, to obtain the final primitive embeddings, $\hat{v}_a$, $\hat{v}_o$ and $\hat{v}_c$. We finally compute the cross entropy losses of the extracted representations with learned embeddings. Besides, we also propose a diversity loss to assist model in learning super-primitives. (b) Illustration of knowledge distillation in t-th. The solid black line represents forward propagation, and the dashed orange line represents back propagation. Notably, we copy new primitive embeddings to previous model and perform knowledge distillation on new categories. All parameters in the previous model will not update with the gradient.

the similar contextuality, where K is smaller than the number of original primitives. Taking the recognition of attribute as an example, by combining the original attribute embedding with the super-object embedding, we can modify the original attribute embedding based on the contextuality provided by super-object embedding. Similarly, the process of leveraging super-attribute to recognize objects follows a similar way.

Actually, when recognizing a single primitive, we cannot know in advance the super-primitive which another primitive in the current composition belongs to. Therefore, we utilize a super-primitive selection module to determine which super-primitive the another primitive should belongs to. First, we can calculate the correlations between the primitive and all the super-primitives in the latent space as follows:

$$w_p = \text{softmax}(f_1(v_p) \cdot f_2(S_p)^T), \tag{4}$$

where $f_1(\cdot)$ and $f_2(\cdot)$ are the projection layers which project them into a latent space. Subsequently, we utilize the correlation coefficient w_p to obtain the final super-primitive:

$$s_p = w_p \cdot S_p \tag{5}$$

However, when the module is equipped with the ability to select among K super-primitives, the model could potentially learn to use only a subset of super-primitives for prediction. Inspired by [Huynh and Elhamifar, 2020], we introduce a loss function $\mathcal{L}_{div}$ to ensure that each of the K super-primitive will be used for prediction:

$$\mathcal{L}_{div} = \sum_k (\sum_i w_k^i)^2, \tag{6}$$

where the term k in the first sum indicates the correlation coefficient between primitive representation and k-th super-primitive. The term i in the second sum pertains the i-th sample in the current batch. This loss encourages the module F to distribute its focus across all super-primitives.

With the assistance of super-primitives, we can reformulate the calculation of similarity scores for primitives as follows:

$$s(v_a, (e_a, s_o)) = \frac{v_a^T}{||v_a||} \cdot \frac{\varphi_a(e_a, s_o)}{||\varphi_a(e_a, s_o)||}, \tag{7}$$

$$s(v_o, (e_o, s_a)) = \frac{v_o^T}{||v_o||} \cdot \frac{\varphi_o(e_o, s_a)}{||\varphi_o(e_o, s_a)||}, \tag{8}$$

where s_a and s_o represent the selected super-primitives. $\varphi_a(\cdot)$ and $\varphi_o(\cdot)$ are the transformation networks for each primitive, which transform the original primitive embedding and the corresponding super-primitive embedding into the final primitive embedding with the information of contextuality.

To learn super-primitives, we utilize cross-entropy loss to guide the transformed primitive embeddings towards the corresponding primitive representation:

$$\mathcal{L}_a = -\sum_{a \in \mathcal{A}^i_{tr}} \log \frac{\exp(s(v_a, (e_a, s_o))/\tau)}{\sum_{a' \in \mathcal{A}^i_{tr}} \exp(s(v_a, (e_{a'}, s_o))/\tau)}, \quad (9)$$

$$\mathcal{L}_o = -\sum_{o \in \mathcal{O}^i_{tr}} \log \frac{\exp(s(v_o, (e_o, s_a))/\tau)}{\sum_{o' \in \mathcal{O}^i_{tr}} \exp(s(v_o, (e_{o'}, s_a))/\tau)}. \quad (10)$$

Additionally, we deploy the cross-entropy loss for recognizing composition:

$$\mathcal{L}_c = -\sum_{(a,o) \in \mathcal{C}^i_{tr}} \log \frac{\exp(s(v_c, (e_a, e_o))/\tau)}{\sum_{(a',o') \in \mathcal{C}^i_{tr}} \exp(s(v_c, (e_{a'}, e_{o'}))/\tau)}, \quad (11)$$

where τ is the temperature. $\mathcal{A}^i_{tr}$, $\mathcal{O}^i_{tr}$ and $\mathcal{C}^i_{tr}$ denote the label space of attribute, object and composition in the training set from the i-th session respectively. Notably, $\mathcal{C}^i_{tr}$ only includes new compositions, while $\mathcal{A}^i_{tr}$ and $\mathcal{O}^i_{tr}$ encompass all the primitives learned so far. Although jointly optimizing all categories is often associated with error propagation, it may even bring a slight improvement in this issue.

3.4 Mitigating Catastrophic Forgetting by Dual Knowledge Distillation

We now introduce the first application of knowledge distillation (KD) in continual learning, namely Learning without Forgetting (LwF) [Li and Hoiem, 2017]. In brief, LwF takes the model trained from the previous sessions to generate soft labels exclusively for the old categories of the new data, and these soft labels are then used as supervision for the current model. The LwF loss is defined as follows:

$$\mathcal{L}_{LwF} = \sum_{n=1}^{N^{t-1}} \pi^{t-1}_n(x) \log \pi^t_n(x), \quad (12)$$

where N^{t-1} indicates all the categories before t-th session. π^{t-1} and π^t are the predictions made by the previous model and the current model.

An illustration of KD between the model trained in previous sessions and the model trained in current session is shown in Fig.2b. In this paper, we propose a dual knowledge distillation which simultaneously mitigates catastrophic forgetting and overfitting. Different from the LwF loss, we also apply KD on the new categories. The reason lies in that, after training on new data, the parameters of main components tend to bias towards the new categories. The dual KD can help the model preserves old primitive embeddings while avoiding the bias of old components towards the new categories. Here, we introduce more details about the dual KD loss:

$$\mathcal{L}_{KD} = \sum_{n=1}^{N^t} \hat{y}_n \log p_n, \quad (13)$$

where N^t indicates the number of all categories up to t-th current session, which can be attributes, objects or compositions. p_n denotes the prediction of the n-th category generated by the model trained in the current session, and $\hat{y}_n$ represents the soft labels generated from the previous model:

$$\hat{y}_n = \begin{cases} \delta(s(V_{t-1}, T_{t-1}(P_{t-1}))), & 1 \leq n \leq N^{t-1} \\ \\ \delta(s(V_{t-1}, T_{t-1}(P_t))), & N^{t-1} < n \leq N^t \end{cases}, \quad (14)$$

where $\delta(\cdot)$ represents softmax operator. For convenience, we simplify the process of obtaining soft labels from previous model into three components $V_{t-1}(\cdot)$, $T_{t-1}(\cdot)$ and P_{t-1}, as shown in the Figure 2b. The subscript $t-1$ indicates that the component comes from the model trained after the $(t-1)$-th session. It is worth noting that when generating soft labels for new categories, we use the primitive embeddings in the current model and keep updating the new primitive embeddings.

Algorithm 1 The optimization procedure of CCZSL.

```
Input: training sequence $\{\mathcal{D}^0_{tr}, \mathcal{D}^1_{tr}, \ldots, \mathcal{D}^T_{tr}\}$,
    learning rate $\gamma$, weight $\lambda$
Output: $\theta = \{\theta_E, \theta_F, \theta_\varphi, \theta_S, \theta_P\}$
randomly initialize parameters $\theta$
for each session $t \in [0, T]$ do
    if $t > 0$ then
        copy old parameters as $\theta^{old}$
        add new primitive embeddings to $\theta_P$
        while not converaged do
            copy new primitive embeddings from $\theta_P$ to $\theta^{old}_P$
            $\theta = \theta - \gamma \nabla_\theta(\mathcal{L}_{CZSL}(\mathcal{D}^t_{tr}; \theta) +$
            $\lambda \mathcal{L}_{KD}(\mathcal{D}^t_{tr}; \theta^{old}, \theta))$
        end while
    else
        while not converaged do
            $\theta = \theta - \gamma \nabla_\theta \mathcal{L}_{CZSL}(\mathcal{D}^i_{tr}; \theta)$
        end while
    end if
end for
return $\theta = \{\theta_E, \theta_F, \theta_\varphi, \theta_S, \theta_P\}$
```

3.5 Training and Inference

At the training phase, we formulate our final loss as follows:

$$\mathcal{L}_{CZSL} = (\mathcal{L}_a + \mathcal{L}_o) \, * \, 0.5 + \mathcal{L}_c + \mathcal{L}_{div}, \quad (15)$$

$$\mathcal{L} = \mathcal{L}_{CZSL} + \lambda \mathcal{L}_{KD}, \quad (16)$$

During the inference phase, we gather the predictions of primitives and compositions to obtain the final prediction:

$$c = \arg\max_{c \in \mathcal{C}} \{ s(v_a, (e_a, s_o)) \cdot s(v_o, (e_o, s_a)) + s(v_c, (e_a, e_o)) \}. \quad (17)$$

4 Experiments

4.1 Experiment Setting

Datasets. We conduct experiments on two widely adopted datasets in CZSL, which are UT-Zappos [Yu and Grauman, 2014] and C-GQA [Naeem *et al.*, 2021]. UT-Zappos is a fine-grained shoes dataset which contains 50,025 images, with 16

Method	**UT-Zappos** (Session Number)					**C-GQA** (Session Number)							
	0	1	2	Avg	Final	0	1	2	3	4	5	Avg	Final
AoP	42.21	20.34	16.25	26.27	+4.19	2.52	1.32	0.97	0.50	0.34	0.27	0.99	+2.30
SymNet	41.93	13.58	9.46	21.66	+8.80	3.42	2.28	1.15	0.65	0.63	0.52	1.44	+1.85
VisProdNN	43.59	17.65	2.78	21.34	+9.12	4.18	0.40	1.19	0.44	0.24	0.15	1.10	+2.19
SCEN	41.39	15.61	8.74	21.91	+8.55	3.43	0.64	0.75	0.26	0.11	0.11	0.88	+2.41
CANet	45.48	19.89	5.05	23.47	+6.99	**5.17**	2.49	1.90	1.17	1.27	1.00	2.17	+1.12
CCZSL	**47.70**	**24.73**	**18.96**	**30.46**		5.07	**3.89**	**3.88**	**2.74**	**2.32**	**1.83**	**3.29**	

Table 1: Performance comparisons with state-of-the-art CZSL methods across UT-Zappos and C-GQA datasets. We reported the AUC in each sessions, the average AUC across all the sessions and final improvement compared to other methods.

Dataset	**Session**	**attr**	**obj**	**tr**	**vl**	**ts**
UT-Zappos	0	8	6	24	7	9
	1	4	3	27	10	14
	2	4	3	32	13	13
C-GQA	0	233	363	2392	958	730
	1	35	58	491	168	172
	2	32	67	772	358	265
	3	36	64	836	366	275
	4	39	62	562	225	166
	5	38	60	539	217	203

Table 2: Session splitting of UT-Zappos and C-GQA.

attributes and 12 objects. It comprises 83 seen compositions and 15/18 (validation/test) unseen compositions. C-GQA is a natural image dataset which contains 39,298 images, with 413 attributes and 764 objects. It includes 5,592 seen compositions and 1,040/923 (validation/test) unseen compositions.

Datasets Split In our CCZSL setting, we assume that introducing new primitives into each session is a prerequisite, followed by the spontaneous introduction of new compositions through the combination of new and old primitives. Whether the composition is seen or unseen follows the same setting as the original dataset. Specifically, We split UT-Zappos into 3 sessions and C-GQA into 6 sessions. Details of splits are present in Table 2, including number of new attributes, new objects, new training compositions, new validation compositions and new testing compositions in each session. We allocate more primitives for the first session, aiming to let the model learn a good initialization.

Evaluation Metrics The harmonic mean H between the top-1 accuracy on seen S and unseen U classes is a widely used evaluation protocol in GZSL, $i.e.$, $H = 2 \times S \times U/(S + U)$. However, due to the model only accessing seen data during training, there is a bias towards seen data during testing, leading to a decrease in the harmonic mean. To alleviate this issue, adding a scalar bias to the prediction of unseen compositions to calibrate the final result has become a widely used technique in CZSL. In this paper, we use the area under the curve (AUC) of H with the scalar bias range from $-\infty$ to ∞ to assess the CZSL performance. And we utilize the average AUC across all the sessions to report the overall performance.

Implementation Details We use ResNet-18 pre-trained on ImageNet to extract 512 dimensional vector, following preceding work [Wang *et al.*, 2023]. The three encoder networks and transformation networks share the same structure of two Fully Connected (FC) layers with ReLU, LayerNorm and Dropout. The super-primitive selection module is implemented with multi-head attention mechanism. The number of super-primitives K is set to 4 for UT-Zappos and 20 for C-GQA. The original primitive embeddings are learned from scratch. For all datasets, we train the model using Adam optimizer with a learning rate of 5×10^{-5}. The temperature factor is 0.05 for all datasets.

4.2 Comparing with the State of the Art

Baseline. Since CCZSL is a newly proposed setting in CZSL field, there does not exist any prior works that can be used for comparison directly. Therefore, we conduct experiments and compare our method with the following state-of-the-art methods in CZSL: 1) AttrAsOp [Nagarajan and Grauman, 2018] views attributes as transformation operations rather than high-dimensional embeddings. It employs various regularization to optimize these operations. 2) SymNet [Li *et al.*, 2020] introduces symmetry into the attribute-object transformation and learns the transformation under the supervision of group theory. 3) VisProdNN [Karthik *et al.*, 2021] revisits visual-only methods and considers CZSL as a multi-task problem, predicting objects and attributes separately. 4) SCEN [Li *et al.*, 2022] learns embedding of attributes and objects in a siamese contrastive space and proposes a State Transition Module to increase the diversity of training compositions. 4) CANet [Wang *et al.*, 2023] proposes to recognize attribute conditioned on objects and designs an attribute hyper learner to generate flexible attribute embeddings.

Results and Analysis. Table 1 summarizes the results of all the comparison methods. When comparing the results on UT-Zappos, our method achieves the best AUC in each session. Specifically, we obtain the highest AUC of 45.48%, 24.73% and 18.96% in each session, surpassing the second-best methods by 2.22%, 4.39% and 2.71% respectively. For the average AUC over all the sessions, we gain 4.19% improvement over the second-best method. Comparing the results on the more challenging dataset C-GQA, we achieve the best AUC in the subsequent incremental sessions. Specifically, our model achieves the best AUC of 3.89%, 3.88%,

Method	UT-Zappos (Session Number)				C-GQA (Session Number)						
	0	1	2	Avg	0	1	2	3	4	5	Avg
Baseline	46.23	19.86	7.33	24.47	4.74	2.58	1.70	1.10	1.16	0.92	2.03
$+\mathcal{L}_{KD}$	45.82	23.71	16.54	28.69	4.78	3.79	3.61	2.72	2.36	1.86	3.19
+Super-Primitive	44.98	24.26	19.37	29.56	4.85	3.95	3.79	2.79	2.43	1.75	3.26
$+\mathcal{L}_{Div}$	47.70	24.73	18.96	30.46	5.07	3.89	3.88	2.74	2.32	1.83	3.29

Table 3: Ablation analysis on the main contributions of our method.

2.74%, 3.32% and 1.83% in each subsequent incremental sessions respectively. At the same time, we gain improvements of 1.40%, 1.98%, 1.57%, 1.05% and 0.83% compared to the second-best methods. For the average AUC over all the sessions, we gain a 1.12% improvement over the second-best method. We attribute the success of the proposed approach to the super-primitives and the dual KD. The improvement in the first session illustrates the effectiveness of our super-primitives. By introducing contextuality into invariant embeddings, we achieve improved recognition of diverse primitives which can further improve the accuracy of predicting compositions. Subsequent improvements highlight the importance of mitigating catastrophic forgetting in this setting. Our proposed dual KD can not only preserve old primitive embeddings, but also maintain the characteristics of old network which can significantly improve the performance.

4.3 Ablation Analysis

Impact of main components. To validate the effectiveness of our method, we conduct ablation analysis in Table 3. Specifically, Baseline is a plain pipeline mentioned in Section 3.2. Then we add the dual KD loss $\mathcal{L}_{KD}$ to mitigate catastrophic forgetting. As we can see, dual KD can effectively mitigate catastrophic forgetting. To help the model better understand the contextuality, we introduce super-primitive into the recognition of primitive which improves the average AUC on both datasets. At last, we add diversity loss $\mathcal{L}_{div}$ to avoid the model from finding shortcuts, which can further improve overall performance.

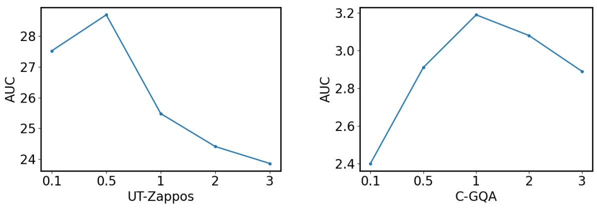

Figure 3: Impact of λ on UT-Zappos and C-GQA.

Impact of λ. The scale of λ can control the balance between learning new knowledge and keeping old knowledge. We can observe that the best average AUCs for both UT-Zappos and C-GQA are achieved when λ equals 0.5 and 1, respectively. This is mainly because C-GQA contains much more and diverse compositions. The significant differences between new and old compositions results in the need for more constraints on old knowledge in C-GQA.

	UT-Zappos	C-GQA
Original KD	27.52	2.56
Our KD	**28.69**	**3.19**

Table 4: Impact of dual knowledge distillation.

Impact of dual knowledge distillation. We conduct experiments to examining the effects of original KD and dual KD. In Table 4, we report the average AUC across all the sessions. We can observe that dual KD can better prevent the performance degradation caused by the model excessively biasing towards the new categories. But the original KD can only maintain the knowledge of old primitive embeddings.

Primitives	Compositions	UT-Zappos	C-GQA
×	×	23.41	2.00
✓	×	**24.47**	**2.03**
✓	✓	21.96	1.53
×	✓	21.05	1.51

Table 5: Impact of joint optimization.

Impact of joint optimization. We report the average AUC of joint optimizing all primitives or compositions. ✓ indicates the use of joint optimization for the corresponding category. We can observe that, jointly optimizing old and new primitives may enhance discrimination between primitives, resulting in improved performance. However, the number of compositions is greater than the number of primitives and the inter-compositions differences are smaller, which will lead to error propagation during joint optimization process.

5 Conclusion

In this paper, we propose a practical and challenging setting called CCZSL, which requires the model to recognize unseen compositions composed of learned primitive set while continually increasing the size of learned primitive set. To solve contextuality issue, we propose to learn super-primitives to capture the similar contextuality in part of compositions, which is beneficial for recognizing the various primitives. Then, we propose a dual knowledge distillation loss to better preserve old knowledge by performing knowledge distillation on new categories. Extensive experiments demonstrate the superiority of our proposed method.

Acknowledgments

This work was supported by the Beijing Natural Science Foundation (No.4242046), the Tangshan Municipal Science and Technology Plan Project (No.23130225E) and the Fundamental Research Funds for the Central Universities (No.2022JBZY019).

References

[Atzmon *et al.*, 2016] Yuval Atzmon, Jonathan Berant, Vahid Kezami, Amir Globerson, and Gal Chechik. Learning to generalize to new compositions in image understanding. *arXiv preprint arXiv:1608.07639*, 2016.

[Chi *et al.*, 2022] Zhixiang Chi, Li Gu, Huan Liu, Yang Wang, Yuanhao Yu, and Jin Tang. Metafscil: A meta-learning approach for few-shot class incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14166–14175, June 2022.

[Dhar *et al.*, 2019] Prithviraj Dhar, Rajat Vikram Singh, Kuan-Chuan Peng, Ziyan Wu, and Rama Chellappa. Learning without memorizing. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 5138–5146, 2019.

[Gautam *et al.*, 2020] Chandan Gautam, Sethupathy Parameswaran, Ashish Mishra, and Suresh Sundaram. Generalized continual zero-shot learning. *arXiv preprint arXiv:2011.08508*, 2020.

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

[Hinton *et al.*, 2015] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[Hu and Wang, 2023] Xiaoming Hu and Zilei Wang. Leveraging sub-class discimination for compositional zero-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 890–898, 2023.

[Huynh and Elhamifar, 2020] Dat Huynh and Ehsan Elhamifar. A shared multi-attention framework for multi-label zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8776–8786, 2020.

[Karthik *et al.*, 2021] Shyamgopal Karthik, Massimiliano Mancini, and Zeynep Akata. Revisiting visual product for compositional zero-shot learning. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021.

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

[Kirkpatrick *et al.*, 2017] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

[Lake *et al.*, 2017] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40:e253, 2017.

[Li and Hoiem, 2017] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2017.

[Li *et al.*, 2020] Yong-Lu Li, Yue Xu, Xiaohan Mao, and Cewu Lu. Symmetry and group in attribute-object compositions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11316–11325, 2020.

[Li *et al.*, 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*, pages 9326–9335, 2022.

[Li *et al.*, 2023] Zechao Li, Hao Tang, Zhimao Peng, Guo-Jun Qi, and Jinhui Tang. Knowledge-guided semantic transfer network for few-shot image recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[Masana *et al.*, 2022] Marc Masana, Xialei Liu, Bartłomiej Twardowski, Mikel Menta, Andrew D Bagdanov, and Joost Van De Weijer. Class-incremental learning: survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(5):5513–5533, 2022.

[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/CVF Conference on Computer Vision and Pattern Recognition*, pages 1792–1801, 2017.

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

[Nagarajan and Grauman, 2018] Tushar Nagarajan and Kristen Grauman. Attributes as operators: factorizing unseen attribute-object compositions. In *Proceedings of the Euro-*

pean Conference on Computer Vision (ECCV), pages 169–185, 2018.

[Palatucci *et al.*, 2009] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. Zero-shot learning with semantic output codes. *Advances in Neural Information Processing Systems*, 22, 2009.

[Saini *et al.*, 2022] Nirat Saini, Khoi Pham, and Abhinav Shrivastava. Disentangling visual embeddings for attributes and objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13658–13667, 2022.

[Skorokhodov and Elhoseiny, 2020] Ivan Skorokhodov and Mohamed Elhoseiny. Class normalization for (continual)? generalized zero-shot learning. *arXiv preprint arXiv:2006.11328*, 2020.

[Tang *et al.*, 2020] Hao Tang, Zechao Li, Zhimao Peng, and Jinhui Tang. Blockmix: meta regularization and self-calibrated inference for metric-based meta-learning. In *Proceedings of the 28th ACM international conference on multimedia*, pages 610–618, 2020.

[Tang *et al.*, 2022] Hao Tang, Chengcheng Yuan, Zechao Li, and Jinhui Tang. Learning attention-guided pyramidal features for few-shot fine-grained recognition. *Pattern Recognition*, 130:108792, 2022.

[Titsias *et al.*, 2019] Michalis K Titsias, Jonathan Schwarz, Alexander G de G Matthews, Razvan Pascanu, and Yee Whye Teh. Functional regularisation for continual learning with gaussian processes. *arXiv preprint arXiv:1901.11356*, 2019.

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

[Wei *et al.*, 2020] Kun Wei, Cheng Deng, Xu Yang, et al. Lifelong zero-shot learning. In *IJCAI*, pages 551–557, 2020.

[Wu *et al.*, 2023a] Yanan Wu, Zhixiang Chi, Yang Wang, and Songhe Feng. Metagcd: Learning to continually learn in generalized category discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1655–1665, October 2023.

[Wu *et al.*, 2023b] Yanan Wu, Tengfei Liang, Songhe Feng, Yi Jin, Gengyu Lyu, Haojun Fei, and Yang Wang. Metazscil: A meta-learning approach for generalized zero-shot class incremental learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10408–10416, 2023.

[Xian *et al.*, 2018] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5542–5551, 2018.

[Yu and Grauman, 2014] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 192–199, 2014.

[Zhang and Feng, 2023] Yang Zhang and Songhe Feng. Enhancing domain-invariant parts for generalized zero-shot learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6283–6291, 2023.

[Zhang *et al.*, 2020] Junting Zhang, Jie Zhang, Shalini Ghosh, Dawei Li, Serafettin Tasci, Larry Heck, Heming Zhang, and C-C Jay Kuo. Class-incremental learning via deep model consolidation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1131–1140, 2020.

[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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 339–355. Springer, 2022.