Learning Conditional Attributes for Compositional Zero-Shot Learning

Qingsheng Wang[1*], Lingqiao Liu[2], Chenchen Jing[3], Hao Chen[3], Guoqiang Liang[1],
Peng Wang[1†], Chunhua Shen[3]

[1]School of Computer Science and Ningbo Institute, Northwestern Polytechnical University, Xi'an, China
[2]School of Computer Science, University of Adelaide, Adelaide, Australia
[3]School of Computer Science, Zhejiang University, Hangzhou, China

Abstract

Compositional Zero-Shot Learning (CZSL) aims to train models to recognize novel compositional concepts based on learned concepts such as attribute-object combinations. One of the challenges is to model attributes interacted with different objects, e.g., the attribute "wet" in "wet apple" and "wet cat" is different. As a solution, we provide analysis and argue that attributes are conditioned on the recognized object and input image and explore learning conditional attribute embeddings by a proposed attribute learning framework containing an attribute hyper learner and an attribute base learner. By encoding conditional attributes, our model enables to generate flexible attribute embeddings for generalization from seen to unseen compositions. Experiments on CZSL benchmarks, including the more challenging C-GQA dataset, demonstrate better performances compared with other state-of-the-art approaches and validate the importance of learning conditional attributes. Code[‡] is available at `https://github.com/wqshmzh/CANet-CZSL`.

1. Introduction

Deep machine learning algorithms today can learn knowledge of concepts to <mark>recognize patterns.</mark> Can a machine compose different learned concepts to generalize to new compositions? Compositional generalization is one of the hallmarks of human intelligence [3, 18]. To make the models equipped with this ability, Compositional Zero-Shot Learning (CZSL) [25] is proposed, where the models are trained to recognize images of unseen compositions composed of seen concepts. In this work, we concentrate on the situation where each composition is composed by attribute (e.g., *wet*) and object (e.g., *apple*). For example, given im-

---
[*]E-mail: wqshmzh@mail.nwpu.edu.cn
[†]Corresponding author. E-mail: peng.wang@nwpu.edu.cn
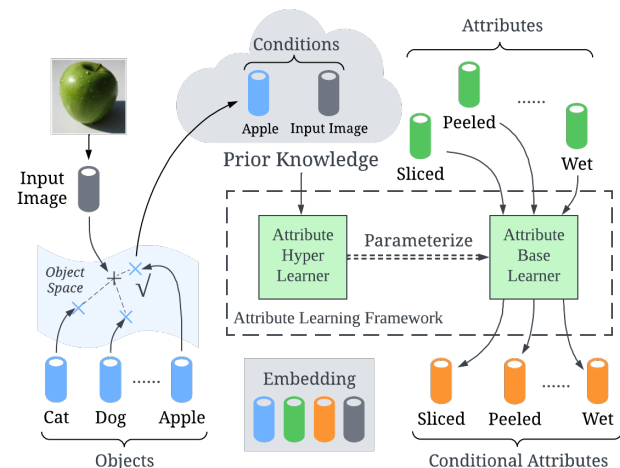[‡]Gitee: `https://gitee.com/wqshmzh/canet-czsl`

Figure 1. The diagram of our work. We aim to learn conditional attributes conditioned on the recognized object and input image through an attribute learning framework containing an attribute hyper learner and an attribute base learner. We first recognize the object in the input image. Then, we feed prior knowledge extracted from the conditions, which are recognized object word embedding and input image visual embedding, to the attribute hyper learner. Finally, conditional attribute embeddings are produced by the attribute base learner parameterized by the attribute hyper learner.

ages of *wet apple* and *dry cat*, a well-trained model can recognize images of new compositions *dry apple* and *wet cat*.

Compositional Zero-Shot Learning of attribute-object compositions requires modeling attributes, objects, and the contextuality between them. Learning to model objects in CZSL is similar to conventional supervised object classification task since the model has access to all objects in CZSL task [33]. Learning to model contextuality between attribute and object is mostly addressed in the literature [23,25,26,31,39–41]. One of the main challenges of CZSL is the appearance diversity of an attribute when composed with different objects, e.g., attribute *wet* in *wet apple* and *wet cat* is different. This reveals that the information of each attribute is dependent on different objects. However, most

recent works in CZSL [4, 27, 32, 33, 42, 45] extract attribute representations irrelevant to the object from seen compositions to infer the unseen compositions. These approaches neglect the nature of attribute diversity and learn concrete attribute representation agnostic to different objects.

In this paper, we learn conditional attributes rather than learning concrete ones in a proposed **C**onditional **A**ttribute **Net**work (CANet). We first conduct analysis to determine the exact conditions by considering the recognition of attribute and object as computing a classification probability of attribute and object conditioned on the input image. By decomposing this probability, we demonstrate that the probability of the input image belonging to an attribute is conditioned on the recognized object and the input image.

We present an attribute learning framework to learn conditional attribute embeddings conditioned on the above two conditions. The framework contains an attribute hyper learner and an attribute base learner, which are sketched in Fig. 1. The attribute hyper learner learns from prior knowledge extracted from the conditions. The attribute base learner is parameterized by the attribute hyper learner and is designed to encode all attribute word embeddings into conditional attribute embeddings. With the attribute learning framework, the attribute embeddings are changed along with the recognized object and input image. Finally, the attribute matching is processed in an attribute space where the input image embedding is projected. The attribute classification logits are computed by cosine similarities between the projected input image embedding and all conditional attribute embeddings. Additionally, we model the contextuality between attribute and object as composing attribute and object word embeddings. We use cosine similarities between the projected input image embedding and all composed attribute-object embeddings to get the classification logits.

Our main contributions are as follows:

- We propose to learn attributes conditioned on the recognized object and input image.
- We propose an attribute learning framework containing an attribute hyper learner and an attribute base learner for learning conditional attribute embeddings.
- Experiments and ablation studies indicate the effectiveness of our proposed conditional attribute network, which further validates the importance of learning conditional attributes in the CZSL task.

2. Related Work

Compositional Zero-Shot Learning. Given descriptions only, we can recognize objects that are never seen before. In conventional Zero-Shot Learning (ZSL), models have access both to images of seen classes and descriptions of seen and unseen classes [19]. In contrast, CZSL presents no description of seen and unseen attribute-object

compositions while all attributes and objects as concepts are seen during training. Recently, works in CZSL are divided into two main streams. One extracts attribute and object words or visual features independently from a composition during training, including learning attributes as linear transformations of objects [27], learning to hierarchically decompose compositions and recompose the concepts with learned visual concepts [41], learning independent prototypes of attributes and objects and compositing prototypes via graph network [32], and learning decomposed prototypes of visual concept features [33] via siamese contrastive embedding network [20]. The other is to learn a compositional space [23], a graph network [2, 26], an episode-based cross-attention module [39], and a contrastive space [1] for contextuality modeling. Also, Yang *et al.* [42] rethink the CZSL task in a decomposable causal way and learn three spaces for attribute, object, and composition classifications. Additionally, with pre-trained large vision language models like CLIP, Nayak *et al.* [30] propose to tune soft prompts as concept embeddings.

Recent work in [12] addresses the problem of attribute diversity. They propose to learn translational attribute features conditionally dependent on the object prototypes. Specifically, they add generic object embedding as the object prototype to the concatenated attribute and object embedding. However, this approach makes the model concentrate more on the composition instead of the attribute, causing the attribute learning degrade to learning the contextuality between attribute and object. On the contrary, we explicitly focus on learning conditional attribute embeddings. The learned conditional attribute embeddings can be changed along with the objects and input images.

Attribute Learning. Learning features of attributes is explored by a large community including image search [16, 34], sentence generation [17], and zero-shot classification [9, 29]. Conventional attribute learning approaches map the attributes into high-dimensional space and train a discriminative classification head without considering the diverse nature of attributes [22, 35, 37]. Our work also learns high-dimensional embeddings to represent attributes. The main difference is that our learned attribute embeddings are conditioned on different objects and input images.

3. Approach

3.1. Task Definition

The task of CZSL aims to learn to classify an image i into a composition c composed by multiple seen concepts, where i and c are unseen during training. Denote sets of images, compositions, attributes, and objects as $\mathcal{I}$, $\mathcal{C}$, $\mathcal{A}$, and $\mathcal{O}$, we have $i \in \mathcal{I}$, $c \in \mathcal{C}$, $a \in \mathcal{A}$, $o \in \mathcal{O}$, and $\mathcal{C} = \mathcal{A} \times \mathcal{O}$. During training, machines have access to seen set $\mathcal{D}_{seen} = \{(i^s, c^s) | i^s \in \mathcal{I}^s, \mathcal{I}^s \subsetneq \mathcal{I}, c \in \mathcal{C}^s, \mathcal{C}^s \subsetneq \mathcal{C}\}$,
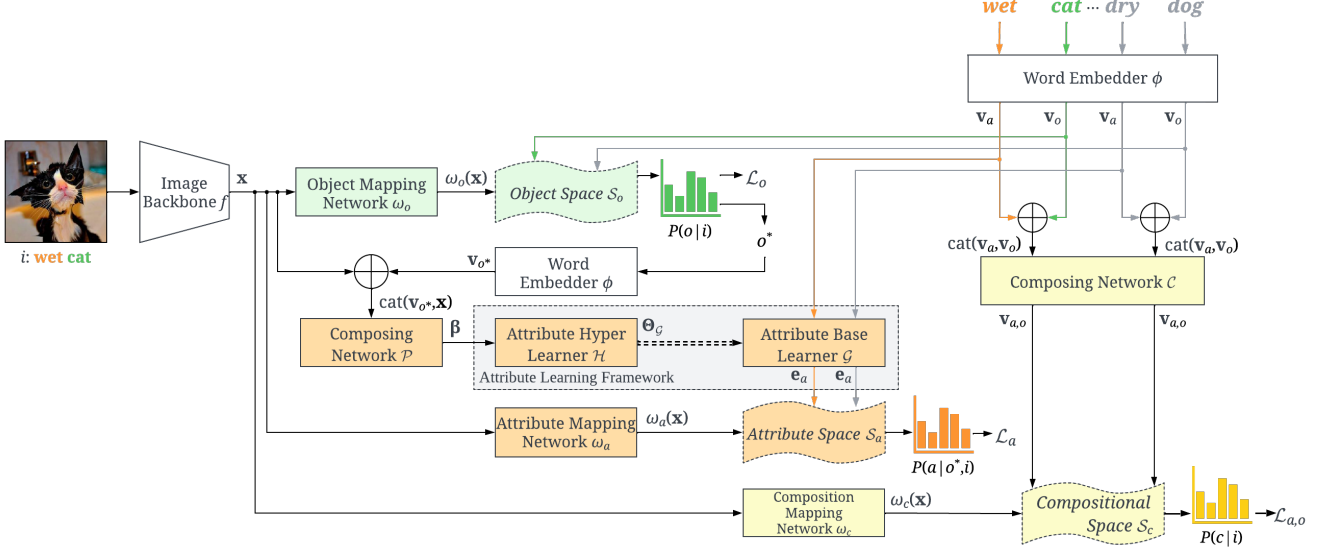
Figure 2. Structure of our proposed CANet. The symbol $\bigoplus$ is channel-wise concatenation. The mapping networks ω_o, ω_a, and ω_c map the input image embedding $\mathbf{x}$ into object, attribute, and composition spaces $\mathcal{S}_o$, $\mathcal{S}_a$, and $\mathcal{S}_c$. All object word embeddings $\mathbf{v}_o$ along with the object-mapped input image embedding are used in $\mathcal{S}_o$ to compute loss $\mathcal{L}_o$ and get the recognized object o^*. The attribute hyper learner $\mathcal{H}$ learns to parameterize the attribute base learner $\mathcal{G}$ using the prior knowledge $\boldsymbol{\beta}$ extracted from the recognized object word embedding $\mathbf{v}_{o^*}$ and $\mathbf{x}$. The conditional attribute embeddings $\mathbf{e}_a$ are encoded by $\mathcal{G}$ parameterized by $\mathcal{H}$. Using all $\mathbf{e}_a$ along with the attribute-mapped input image embedding in space $\mathcal{S}_a$ to compute loss $\mathcal{L}_a$. Loss $\mathcal{L}_{a,o}$ is computed using all compositional word embeddings produced by composing network $\mathcal{C}$ and the composition-mapped input image embedding in space $\mathcal{S}_c$.

attribute set $\mathcal{A}$, and object set $\mathcal{O}$, where $\mathcal{I}^s$ and $\mathcal{C}^s$ are sets of images and compositions seen when training. Also, evaluation of algorithms requires unseen set $\mathcal{D}_{unseen} = \{(i^u, c^u) | i^u \in \mathcal{I}^u, c^u \in \mathcal{C}^u, \mathcal{I}^u \subsetneqq \mathcal{I}, \mathcal{C}^u \subsetneqq \mathcal{C}\}$ used for validation and testing. In conventional ZSL, $\mathcal{I}^s \cap \mathcal{I}^u = \varnothing$, $\mathcal{C}^s \cap \mathcal{C}^u = \varnothing$, i.e. unseen images and compositions are not overlapped with the seen ones. Here, we follow the setting of Generalized Zero-Shot Learning (GZSL) where images in $\mathcal{I}^s$ and $\mathcal{I}^u$ and labels in $\mathcal{C}^s$ and $\mathcal{C}^u$ appear during validation and testing. GZSL is a challenging setting with larger label space and a strong bias of seen compositions to unseen ones, relaxing the less realistic assumption in conventional ZSL that test data only belongs to unseen classes.

3.2. Conditional Attribute Network

Determining Conditions. In CZSL, it is common to learn to classify attributes and objects besides compositions. We first assume that the model recognizes the input image i as attribute a^* and object o^*. The recognizing score can be formulated as a conditional probability $P(a^*, o^*|i)$ conditioned on input image i. We propose to decompose this probability to express the attribute and object recognition as a one-label classification task. According to multi-variable conditional probability formulation, we have:

$$P(a^*, o^*|i) = P(a^*|o^*, i)P(o^*|i) \qquad (1)$$

where $P(o^*|i)$ is probability of image i belonging to object o^* and $P(a^*|o^*, i)$ represents the probability of attribute a^*

conditioned on the joint presentation of o^* and i. This indicates that the recognition of an attribute is conditioned on the recognized object and input image. To better solve the attribute diversity problem, we consider the information of the recognized object and input image as conditions for conditional attribute encoding.

Object Recognition. Object recognition requires learning to map the input image embedding into an object space. To incorporate object semantic information, we compute cosine similarities between object-mapped image embedding $\omega_o(\mathbf{x})$ and all object word embeddings $\mathbf{v}_o$ instead of directly learning a classification head:

$$\langle \omega_o(\mathbf{x}), \mathbf{v}_o \rangle = \frac{\omega_o(\mathbf{x})^\top \mathbf{v}_o}{\|\omega_o(\mathbf{x})\|_2 \|\mathbf{v}_o\|_2} \qquad (2)$$

where $\mathbf{x} = f(i) \in \mathcal{X}$ is the visual embedding of input image i in visual space $\mathcal{X}$ extracted by image backbone f (e.g., ResNet-18 [11]), ω_o is the mapping from $\mathcal{X}$ to object semantic space $\mathcal{S}_o$, $\mathbf{v}_o = \phi(o)$, $\mathbf{v}_o \in \mathcal{S}_o$ is object word embedding in $\mathcal{S}_o$ extracted by word embedder ϕ (e.g., word2vec [24], FastText [7]). The recognized object $o^* = \arg\max_{o \in \mathcal{O}} \langle \omega_o(\mathbf{x}), \phi(o) \rangle$.

Learning Conditional Attributes. With the recognized object o^* and visual embedding $\mathbf{x}$ of input image i, we learn the attribute hyper learner $\mathcal{H}$ and attribute base learner $\mathcal{G}$ in the proposed attribute learning framework to extract attribute embeddings conditioned on the recognized object o^*

and input image i. We consider the information of o^* and i as prior knowledge for $\mathcal{H}$. Specifically, the prior knowledge is implemented as a feature vector $\boldsymbol{\beta}$:

$$\boldsymbol{\beta} = \mathcal{P}(\text{cat}(\mathbf{v}_{o^*}, \mathbf{x})) \tag{3}$$

where $\mathbf{v}_{o^*} = \phi(o^*)$ is word embedding of o^*, $\mathcal{P}$ is a composing network for $\mathbf{v}_{o^*}$ and $\mathbf{x}$, $\text{cat}(\cdot)$ is channel-wise concatenation operation. With the prior knowledge, the attribute hyper learner $\mathcal{H}$ can parameterize $\mathcal{G}$ as:

$$\boldsymbol{\Theta}_{\mathcal{G}} = \mathcal{H}(\boldsymbol{\beta}; \boldsymbol{\Theta}_{\mathcal{H}}) \tag{4}$$

where $\boldsymbol{\Theta}_{\mathcal{H}}$ is the set of randomly initialized parameters of $\mathcal{H}$, $\boldsymbol{\Theta}_{\mathcal{G}}$ is the set of generated parameters of $\mathcal{G}$. In this way, the attribute embeddings $\mathbf{e}_a$ conditioned on o^* and i can be encoded via $\mathcal{G}$ parameterized by $\mathcal{H}$:

$$\mathbf{e}_a = \mathcal{G}(\mathbf{v}_a; \boldsymbol{\Theta}_{\mathcal{G}}) \tag{5}$$

where $\mathbf{v}_a = \phi(a)$ is word embedding of attribute word a.

Modeling Contextuality. Although the conditional attribute embeddings are related to the recognized object and input image, object embeddings are also supposed to be influenced by attributes. Therefore, we model the contextuality of attribute-object compositions to address the relationships between them. We follow the work of Mancini *et al.* [23] in their closed world setting that contextuality is modeled as the mixture of attribute and object word embeddings to extract attribute-object compositional embeddings:

$$\mathbf{v}_{a,o} = \mathcal{C}(\text{cat}(\mathbf{v}_a, \mathbf{v}_o)) \tag{6}$$

where $\mathcal{C}$ is a composing network for $\mathbf{v}_a$ and $\mathbf{v}_o$.

The entire structure of our model is shown in Fig. 2.

3.3. Training Objectives

Similar to object recognition, attribute or composition recognition is also implemented by computing cosine similarities $\langle \omega_a(\mathbf{x}), \mathbf{v}_a \rangle$ or $\langle \omega_c(\mathbf{x}), \mathbf{v}_{a,o} \rangle$ between attribute-mapped or composition-mapped image embeddings, *i.e.*, $\omega_a(\mathbf{x})$ or $\omega_c(\mathbf{x})$, and attribute word embeddings or attribute-object compositional embeddings, *i.e.*, $\mathbf{v}_a$ or $\mathbf{v}_{a,o}$:

$$\langle \omega_a(\mathbf{x}), \mathbf{e}_a \rangle = \frac{\omega_a(\mathbf{x})^\top \mathbf{e}_a}{\|\omega_a(\mathbf{x})\|_2 \|\mathbf{e}_a\|_2} \tag{7}$$

$$\langle \omega_c(\mathbf{x}), \mathbf{v}_{a,o} \rangle = \frac{\omega_c(\mathbf{x})^\top \mathbf{v}_{a,o}}{\|\omega_c(\mathbf{x})\|_2 \|\mathbf{v}_{a,o}\|_2} \tag{8}$$

The recognition probability $P(o^*|i)$, $P(a^*|o^*, i)$, and $P(c^*|i)$ are normalized cosine similarities, where $P(c^*|i)$ is the probability of input image i belonging to the recognized attribute-object composition c^*. As shown in Fig. 2, our model learns three embedding spaces: attribute space $\mathcal{S}_a$, object space $\mathcal{S}_o$, and attribute-object compositional space $\mathcal{S}_c$. Therefore, we incorporate three separate cross-entropy losses to maximize the three recognition probabilities to make the model optimized in these three spaces. The losses are as follows:

$$\mathcal{L}_a = -\sum_{a \in \mathcal{A}} \log \frac{\exp(\langle \omega_a(\mathbf{x}), \mathbf{e}_a \rangle / \tau)}{\sum_{a' \in \mathcal{A}} \exp(\langle \omega_a(\mathbf{x}), \mathbf{e}_{a'} \rangle / \tau)} \tag{9}$$

$$\mathcal{L}_o = -\sum_{o \in \mathcal{O}} \log \frac{\exp(\langle \omega_o(\mathbf{x}), \mathbf{v}_o \rangle / \tau)}{\sum_{o' \in \mathcal{O}} \exp(\langle \omega_o(\mathbf{x}), \mathbf{v}'_o \rangle / \tau)} \tag{10}$$

Also, for composition recognition, we have:

$$\mathcal{L}_{a,o} = -\sum_{(a,o) \in \mathcal{C}} \log \frac{\exp(\langle \omega_c(\mathbf{x}), \mathbf{v}_{a,o} \rangle / \tau)}{\sum_{(a',o') \in \mathcal{C}} \exp(\langle \omega_c(\mathbf{x}), \mathbf{v}_{a',o'} \rangle / \tau)} \tag{11}$$

where τ is temperature factor [46]. Finally, the training loss as a whole linearly combines the three losses above:

$$\mathcal{L} = \frac{\mathcal{L}_a + \mathcal{L}_o}{2} + \mathcal{L}_{a,o} \tag{12}$$

3.4. Inference

During validation and testing, we incorporate a linear normalization function g for cosine similarities:

$$g(d) = (1 + d) * 0.5 \tag{13}$$

Then, we have $P(a|o^*, i) = g(\langle \omega_a(\mathbf{x}), \mathbf{e}_a \rangle)$, $P(o|i) = g(\langle \omega_o(\mathbf{x}), \mathbf{v}_o \rangle)$, and $P(c|i) = g(\langle \omega_c(\mathbf{x}), \mathbf{v}_{a,o} \rangle)$. The inference rule is parameterized as:

$$s = (1 - \alpha)P(c|i) + \alpha P(a|o^*, i)P(o|i) \tag{14}$$

where α is the weight factor controlling balance.

4. Experiments

In this section, experiments are conducted following the concrete introductions of datasets, metrics, implementation details, and baselines. Then, we report ablation results to demonstrate the effectiveness of our model.

4.1. Experimental Setup

Datasets We conduct experiments with three widely adopted datasets in the CZSL task, which are MIT-States [14], UT-Zappos50K [43, 44], and C-GQA [26]. MIT-States contains 53753 crawled web images labeled with 1962 attribute-object (e.g., *mossy highway*) compositions. This dataset has 30338, 10420, and 12995 training, validation, and testing images [31] labeled with 1262, 600, and 800 compositions. In validation and test sets, the numbers of seen and unseen compositions are the same. All compositions are composed of 115 attributes and 245 objects. UT-Zappos50K is made up of 50025 images labeled with 116 fine-grained shoe classes composed of 16 attributes (e.g., *rubber*) and 12 objects (e.g., *sneaker*). This dataset has 22998, 3214, and 2914 training, validation, and testing images [31] labeled with 83, 30, and 36 compositions. Also, numbers of seen and unseen compositions in validation and test sets share the same quantity. C-GQA is created based on Stanford GQA dataset [13] used for VQA task. C-GQA contains 39298 images labeled with over 9.5k compositions composed of 453 attributes and 870 objects. This dataset has 26920, 7280, and 5098 training, validation, and testing images labeled with 6963, 2541, and 2069 compositions. Detailed splits are presented in Tab. 1.

Dataset		Training			Validation			Test		
	$\mathcal{A}$	$\mathcal{O}$	$\mathcal{C}_s$	$\mathcal{I}$	$\mathcal{C}_s$	$\mathcal{C}_u$	$\mathcal{I}$	$\mathcal{C}_s$	$\mathcal{C}_u$	$\mathcal{I}$
UT-Zappos50K [43, 44]	16	12	83	22998	15	15	3214	18	18	2914
MIT-States [14]	115	245	1262	30338	300	300	10420	400	400	12995
C-GQA [26]	453	870	6963	26920	1173	1368	7280	1022	1047	5098

Table 1. Detailed dataset splits of UT-Zappos50K, MIT-States, and C-GQA in training, validation, and test sets.

Metrics To demonstrate the advances in attribute learning, we report the attribute and object classification accuracies (best attr and best obj). The setting of GZSL requires both seen and unseen compositions to exist during validation and testing. As a result, there is an inherent bias of seen against unseen compositions. We follow the evaluation protocols proposed in [8] where a scalar bias is added to final activations of classes of seen compositions to calibrate the model. As the scalar varies from negative infinity to positive infinity, there must be a best operating point at which the bias between seen and unseen compositions is the lowest. We report the results in terms of the best accuracies of seen images (best seen), unseen images (best unseen), best Harmonic Mean (best HM), and Area Under Curve (AUC) with different scalar biases.

Implementations We consider image backbone f as ResNet-18 pre-trained on ImageNet [10] to extract 512 dimension vectors following preceding works. The mapping networks ω_a, ω_o, and ω_c share the similar structure of two Fully Connected (FC) layers with ReLU [28], LayerNorm [5], and Dropout [36] following the first FC layer. We adopt word embedder ϕ as 600-dimensional word2vec+FastText for MIT-States, 300-dimensional Fast-Text for UT-Zappos50K, and word2vec for C-GQA. The layer structures of $\mathcal{G}$, $\mathcal{P}$, and $\mathcal{C}$ are the same as the mapping networks, where ReLU is added in $\mathcal{P}$ and $\mathcal{C}$ to the last FC layer. Weight generation for the attribute base learner $\mathcal{G}$ through the attribute hyper learner $\mathcal{H}$ requires more parameters and makes learning difficult, as noted by Bertinetto *et al.* [6]. Therefore, we adopt weight factorization in [38] to reduce parameters for the attribute hyper learner $\mathcal{H}$, that is

$$\mathbf{e}_a^{(i)} = (\mathbf{v}_a \mathbf{W}_{\mathcal{G}}^{(i)} + \mathbf{b}_{\mathcal{G}}^{(i)}) \odot \lambda_{\mathcal{G}}^{(i)} \qquad (15)$$

$$\lambda_{\mathcal{G}}^{(i)} = \mathcal{H}(\boldsymbol{\beta}; \boldsymbol{\Theta}_{\mathcal{H}}^{(i)}) \qquad (16)$$

where (i) indicates ith FC layer, $\mathbf{W}_{\mathcal{G}}^{(i)}$ and $\mathbf{b}_{\mathcal{G}}^{(i)}$ are the weight matrix and bias vector of ith FC layer in $\mathcal{G}$, $\{\mathbf{W}_{\mathcal{G}}^{(i)}, \mathbf{b}_{\mathcal{G}}^{(i)}\} \subset \boldsymbol{\Theta}_{\mathcal{G}}$, $\boldsymbol{\Theta}_{\mathcal{H}}^{(i)} \subset \boldsymbol{\Theta}_{\mathcal{H}}$, and $\odot$ denotes element-wise multiplication. During training, we fix the image backbone f and train other modules using Adam [15] optimizer and an Nvidia GeForce GTX 1080Ti GPU with a learning rate and weight decay of 0.00005. The batch size is 256. The temperature factor τ, weight factor α, and the maximum number of epochs are set to 0.02, 0.4, and 500 for UT-Zappos50K; 0.05, 0.2, and 800 for MIT-States; and 0.05, 0.4, and 1000 for C-GQA.

Baselines We conduct experiments with the following algorithms: 1) *AttrAsOp* [27] treats attributes as linear transformations on object vectors instead of data points in some high-dimensional space and optimizes the transformations through several regularizers in the loss function. 2) *TMN* [31] constructs task-driven modular networks in semantic space configured through a gating function conditioned on the task. 3) *SymNet* [21] proposes symmetry property in attribute-object compositions and group axioms as objectives in an end-to-end manner. 4) *CGE*ff [26] exploits dependencies between attributes, objects, and compositions through an end-to-end graph formulation where "ff" means fixed image feature backbone. 5) *CompCos* [23] learns a mapping from image features to semantic space of compositions and computes cosine similarities between them. 6) *DeCa* [42] rethinks the CZSL task in a decomposable causal perspective and learns three independent mappings from image feature space to attribute, object, and composition semantic space. Cosine similarities are also adopted. 7) *SCEN* [20] computes visual prototypes of attributes and objects in a siamese contrastive space and proposes a designed State Transition Module to increase the diversity of training compositions.

4.2. Quantitative Analysis

All results are computed on test sets of all datasets and from their published papers and [26] for a fair comparison. We report quantitative results with the best AUC in Tab. 2.

From Tab.2, our model outperforms other state-of-the-art algorithms in terms of best attr, best unseen, and AUC in all three datasets including the recently proposed C-GQA, indicating the better attribute learning performance and generalization ability from seen to unseen compositions. Specifically, our model performs much better on C-GQA with more state-of-the-art results although it is a much more challenging dataset than MIT-States and UT-Zappos50K.

For UT-Zappos50K, the observations are that our model boosts attribute recognition accuracy, unseen image classification accuracy, and AUC from 47.3%, 63.1%, and 32.0% of SCEN to the new state-of-the-art of 48.4%, 66.3%, and 33.1% with 1.1%, 3.2%, and 1.1% improvement respectively. For MIT-States, our model achieves 30.2%, 32.6%, 26.2%, and 5.4% accuracies for attribute and object classification, unseen image classification, and AUC on the test

Algorithm		UT-Zappos50K						MIT-States						C-GQA					
Name	Venue	Att.	Obj.	S.	U.	HM	AUC	Att.	Obj.	S.	U.	HM	AUC	Att.	Obj.	S.	U.	HM	AUC
AttrAsOp [27]	ECCV'18	38.9	69.6	59.8	54.2	40.8	25.9	21.1	23.6	14.3	17.4	9.9	1.6	8.3	12.5	11.8	3.9	2.9	0.3
TMN [31]	ICCV'19	40.8	69.9	58.7	60.0	45.0	29.3	23.3	26.5	20.2	20.1	13.0	2.9	9.7	20.5	21.6	6.3	7.7	1.1
SymNet [21]	CVPR'20	41.3	68.6	49.8	57.4	40.4	23.4	26.3	28.3	24.4	25.2	16.1	3.0	15.0	23.1	27.0	10.8	10.9	2.2
CGE$_{ff}$ [26]	CVPR'21	45.0	73.9	56.8	63.6	41.2	26.4	27.9	32.0	28.7	25.3	17.2	5.1	12.7	26.9	27.5	11.7	11.9	2.5
CompCos [23]	CVPR'21	44.7	73.5	59.8	62.5	43.5	28.7	27.9	31.8	25.3	24.6	16.4	4.5	-	-	-	-	-	-
DeCa [42]	TMM'22	-	-	62.7	63.1	46.3	31.6	-	-	29.8	25.2	18.2	5.3	-	-	-	-	-	-
SCEN [20]	CVPR'22	47.3	**75.6**	**63.5**	63.1	**47.8**	32.0	28.2	32.2	**29.9**	25.2	**18.4**	5.3	13.6	**27.9**	28.9	12.1	12.4	2.9
Ours		**48.4**	72.6	61.0	**66.3**	47.3	**33.1**	**30.2**	**32.6**	29.0	**26.2**	17.9	**5.4**	**17.5**	22.3	**30.0**	**13.2**	**14.5**	**3.3**

Table 2. Quantitive results on test sets of all datasets with the state-of-the-art in terms of best attr (Att.), best obj (Obj.), best seen (S.), best unseen (U.), best HM (HM), and AUC.

set, providing 2.0%, 0.4%, 1.0%, and 0.1% improvements on the recently proposed SCEN as the new state-of-the-art results. This indicates that the proposed conditional attribute network can truly improve the attribute recognition performance and consequently the unseen image classification and AUC.

For the more challenging dataset C-GQA, since it is significantly harder than MIT-States and UT-Zappos50K with 5.5× and 0.9× composition labels and images in the training set compared with MIT-States, our model outperforms all other algorithms except best obj with 3.9%, 1.1%, 1.1%, 2.1%, and 0.4% boosting in terms of best attr, best seen, best unseen, best HM, and AUC in the testing set. This indicates that the proposed conditional attribute network makes a critical contribution when facing a more challenging dataset even if the object recognition accuracy is lower.

We give an analysis of the importance that attributes should be conditioned on objects. First, note that although DeCa also learns attribute, object, and composition spaces separately, it learns attributes as static embeddings independent from objects, causing lower best unseen and AUC on UT-Zappos50K and MIT-States compared with ours in Tab. 2. Next, different baselines incorporate different techniques to handle the CZSL task though, they learn static attribute embeddings too, producing lower best attr, best unseen, and AUC. Then, the proposed method performs much better on C-GQA compared with other baselines. All the above phenomena demonstrate that attributes should be conditioned on objects and performance on datasets with larger label space can gain more boosts in this way.

From the results of the three datasets above, we observe an interesting phenomenon. Although the results of best obj in three datasets are all lower than that of SCEN, results of best unseen are all higher accompanied by higher results of best attr. We speculate the reason is that some objects have more attributes (or are more dominated) in unseen compositions and the misclassified objects recognized

by our model are less dominated (i.e. have few attributes or are long-tailed in terms of attribute). As a result, with the correctly predicted objects dominated in unseen compositions, the more correctly classified attributes, the higher the results of best unseen.

4.3. Qualitative Analysis

In this section, we present some qualitative results of novel compositions with top-3 predictions on UT-Zappos50K, MIT-States, and C-GQA in Fig. 3. We show results for each dataset in each row. Images whose top prediction matches the label are shown in the first three columns and the rest columns show wrong results. For UT-Zappos50K, the remaining two answers of all images can match at least one label factor. For some instances in MIT-States, we can notice that the top and second predictions can both describe the image. For example, for the image labeled with *winding stream*, there is sunlight reflecting from the stream and creek is the synonym of stream. Therefore, *sunny creek* can also be the label of the image. Another example is that image labeled with *wide valley* also present *cloudy cloud* in the blue sky located in the upper part. As a result, the model has difficulty deciding what to predict. This reflects that labels in MIT-States are heavy in noise. For C-GQA where labels are clear, our model can produce more answers that match the label factors in the remaining two predictions, which indicates the better performance and robustness of our model.

Additionally, we present wrong predictions in the last two columns. It can be noticed that our model can predict correct answers in most top-3 predictions. As for the image of MIT-States labeled as *broken bottle*, our model predicts the attribute as *small* because it is difficult to focus on a certain attribute of the bottle since the bottle in this image is both small and broken. Besides, images in the training set are limited in the status of objects. For example, training images of *sheep* hardly present the action of laying down

Variant		UT-Zappos50K						C-GQA					
#	Name	Att.	Obj.	S.	U.	HM	AUC	Att.	Obj.	S.	U.	HM	AUC
(1)	w/o $\mathcal{L}_a + \mathcal{L}_o$	46.1	**74.3**	61.5	64.7	46.1	31.7	10.8	**30.6**	29.8	12.8	13.4	3.0
(2)	w/o $\mathcal{L}_c$	41.9	60.7	59.5	54.7	45.7	28.3	14.8	17.1	28.1	11.2	12.4	2.6
(3)	w/o $\mathcal{G} + \mathcal{H} + \mathcal{P}$	47.0	74.0	59.9	65.8	46.3	31.7	14.8	27.8	29.9	13.1	14.6	3.1
(4)	w/o $\mathcal{P}$	46.7	73.2	60.7	64.5	47.5	31.6	14.5	26.4	30.1	13.0	14.5	3.1
(5)	w/o $\mathcal{H}$	46.2	70.3	58.5	62.7	46.3	30.5	13.9	25.8	29.1	11.2	12.5	2.4
(6)	w/o $\mathbf{x}$ for $\mathcal{H}$	45.6	71.3	**61.6**	62.8	44.8	30.1	13.4	20.8	30.2	12.7	13.9	2.9
(7)	**Full**	**48.4**	72.6	61.0	**66.3**	**47.3**	**33.1**	**17.5**	22.3	**30.0**	**13.2**	14.5	**3.3**

Table 3. Ablation results on test sets in terms of best attr (Att.), best obj (Obj.), best seen (S.), best unseen (U.), best HM (HM), and AUC.

Figure 3. Qualitative Results. We demonstrate top-3 predictions of some instances using our proposed model.

causing our model mistakenly classify the sheep in the image labeled as *light sheep* into *elephant*. Thus, with the recognized object *elephant*, the model can only focus on attributes conditioned on *elephant* and find the appropriate attribute that matches the image.

4.4. Ablation Study

In this section, we ablate the proposed model to evaluate the performance of each module. The ablation study is conducted on test sets of UT-Zappos50K, C-GQA. To achieve more convincing ablation results, we do not choose MIT-States because images in MIT-States are labeled using automatic search. As a result, the noise of labels is too heavy to be used for evaluations, as is pointed out by Atzmon et.al. [4]. Ablation results are reported in Tab. 3. The detailed ablation process is as follows:

We first study the effects of recognizing compositions only and recognizing attributes and objects without compositions, which are corresponded to variants (1) and (2) in Tab. 3, and report results in terms of six metrics used in Tab. 2. Compared with variant (6), the results of each variant mostly decline, indicating the importance of recognizing attributes, objects, and compositions jointly. As to the results of Obj. in UT-Zappos50K and C-GQA from variants

(1) and (6), we can observe that those object classification accuracies decline. This is mainly because when recognizing objects individually each object presents visual diversity caused by different attributes, e.g., *sliced apple* is different from *apple* in visual appearance since a sliced apple is sliced into multiple pieces. However, this phenomenon does not affect other results, especially for the dataset with larger label space, e.g., C-GQA.

Next, with attributes, objects, and compositions being recognized, we ablate $\mathcal{G} + \mathcal{H} + \mathcal{P}$ (where attribute word embeddings are directly used in $\mathcal{S}_a$), $\mathcal{H}$ (where the concatenation of β and $\mathbf{v}_a$ is fed to $\mathcal{G}$), $\mathcal{P}$ (where the concatenation of $\mathbf{v}_{o*}$ and $\mathbf{x}$ is directly fed to $\mathcal{H}$), and $\mathbf{x}$ in β that is fed to $\mathcal{H}$, which are corresponded to variants (3)-(6) in Tab. 3. Note that if $\mathcal{P}$ is presented then $\mathcal{G}$ is required to be presented also because the concatenation of β and $\mathbf{v}_a$ requires changing the number of dimensions to that of $\mathbf{x}$ through $\mathcal{G}$. Also, $\mathcal{G}$ can not be ablated only otherwise the model changes to variant (3). Compared to variant (7), all results are mostly declined, indicating that optimizing the model with $\mathcal{G}$, $\mathcal{H}$, $\mathcal{P}$, and $\mathbf{x}$ is essential. As for variant (5), it performs worse on seen and unseen images compared with variant (4), indicating the importance of using the attribute hyper learner. Comparing results of variants (5) with (6), we observe that
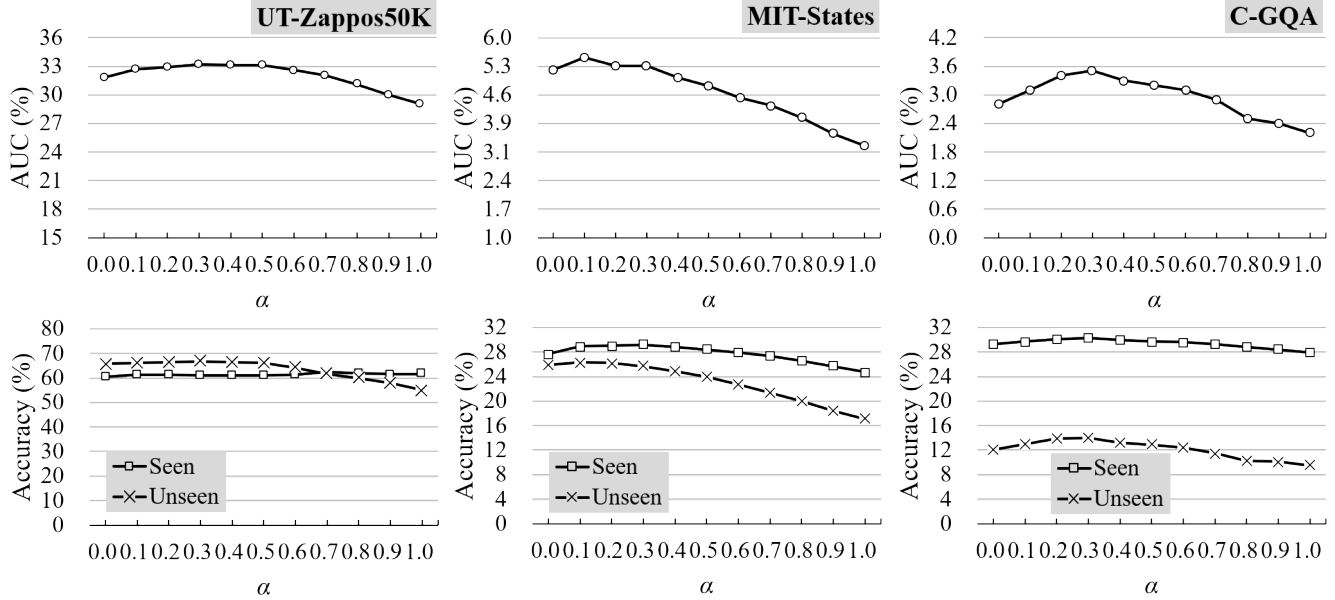
Figure 4. Influence of weighting factor α on AUC, seen accuracy, and unseen accuracy.

adding the attribute hyper learner $\mathcal{H}$ for $\mathcal{G}$ without image visual embedding $\mathbf{x}$ marginally increases overfitting to seen images. Results of variant (6) present the best generalization compared with variants (3)-(5) because adding $\mathbf{x}$ in β for $\mathcal{H}$ provides diverse instances that are seen for $\mathcal{H}$ during training since the number of image visual embeddings is far more than the number of object word embeddings.

Lastly, we conduct experiments to study the impact of weighting factor α in Eq. (14) on all datasets. In detail, we present results of AUC, seen accuracy, and unseen accuracy in Fig. 4 with $\alpha \in [0, 1]$ with an interval of 0.1. It can be observed that all results of AUC increase first and then decrease with α increases from 0.0. The peaks of AUC are reached when α equals 0.5, 0.1, and 0.3 in UT-Zappos50K, MIT-States, and C-GQA. This phenomenon reveals that learning to classify attributes, objects, and compositions all contribute to making our model reach optimal. Additionally, it can be noticed that the results of AUC are all generally declined when α changes from 0.0 to 1.0. This is because attribute-object compositions involve not only the side information of attribute and object, but also the contextuality between attribute and object. As for the results of seen and unseen accuracies reported in the second row, the same trend can be observed. Specifically, results of unseen accuracy are more sensitive than seen accuracy and also have peaks with various α, indicating that α also has an impact on the generalization ability of our model. In conclusion, using a small weighting factor α is always better than using a larger one, indicating that modeling contextuality between attribute and object is a bit more beneficial to the CZSL task than separately classifying attributes and objects.

5. Conclusion

In this work, we address the attribute diversity problem in Compositional Zero-Shot Learning. As a solution, we propose a Conditional Attribute Network (CANet) to learn attributes conditioned on the recognized object and input image. We first decompose the probability of attribute and object recognition conditioned on the input image to lay a foundation for learning conditional attributes. Then, we build an attribute learning framework to encode conditional attribute embeddings. Experiments show that our model outperforms recent CZSL approaches and achieves new state-of-the-art results. Despite the better attribute recognition performance, a limitation is that our model is less qualified to handle object long-tailed distribution in terms of attribute mentioned in Sec. 4.2. Future works can be focused on solving the problem above while learning conditional attributes.

6. Acknowledgements

References

[1] Muhammad Umer Anwaar, Rayyan Ahmad Khan, Zhihui Pan, and Martin Kleinsteuber. A contrastive learning approach for compositional zero-shot learning. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 34–42, 2021. 2

[2] Muhammad Umer Anwaar, Zhihui Pan, and Martin Kleinsteuber. On leveraging variational graph embeddings for open world compositional zero-shot learning. *arXiv preprint arXiv:2204.11848*, 2022. 2

[3] 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. 1

[4] Yuval Atzmon, Felix Kreuk, Uri Shalit, and Gal Chechik. A causal view of compositional zero-shot recognition. *Advances in Neural Information Processing Systems*, 33:1462–1473, 2020. 2, 7

[5] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5

[6] Luca Bertinetto, João F Henriques, Jack Valmadre, Philip Torr, and Andrea Vedaldi. Learning feed-forward one-shot learners. *Advances in neural information processing systems*, 29, 2016. 5

[7] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146, 2017. 3

[8] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *European conference on computer vision*, pages 52–68. Springer, 2016. 5

[9] Hui Chen, Zhixiong Nan, Jingjing Jiang, and Nanning Zheng. Learning to infer unseen attribute-object compositions. *arXiv preprint arXiv:2010.14343*, 2020. 2

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[12] He Huang, Wei Tang, Jiawei Zhang, and Philip S Yu. Translational concept embedding for generalized compositional zero-shot learning. *arXiv preprint arXiv:2112.10871*, 2021. 2

[13] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 4

[14] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1383–1391, 2015. 4, 5

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[16] Adriana Kovashka, Devi Parikh, and Kristen Grauman. Whittlesearch: Image search with relative attribute feedback. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2973–2980. IEEE, 2012. 2

[17] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Babytalk: Understanding and generating simple image descriptions. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2891–2903, 2013. 2

[18] 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, 2017. 1

[19] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE conference on computer vision and pattern recognition*, pages 951–958. IEEE, 2009. 2

[20] 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. 2, 5, 6

[21] 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. 5, 6

[22] Yongxi Lu, Abhishek Kumar, Shuangfei Zhai, Yu Cheng, Tara Javidi, and Rogerio Feris. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5334–5343, 2017. 2

[23] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Open world compositional zero-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5222–5230, 2021. 1, 2, 4, 5, 6

[24] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013. 3

[25] Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1792–1801, 2017. 1

[26] 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. 1, 2, 4, 5, 6

[27] Tushar Nagarajan and Kristen Grauman. Attributes as operators: factorizing unseen attribute-object compositions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 169–185, 2018. 2, 5, 6

[28] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010. 5

[29] Zhixiong Nan, Yang Liu, Nanning Zheng, and Song-Chun Zhu. Recognizing unseen attribute-object pair with generative model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8811–8818, 2019. 2

[30] Nihal V Nayak, Peilin Yu, and Stephen H Bach. Learning to compose soft prompts for compositional zero-shot learning. *arXiv preprint arXiv:2204.03574*, 2022. 2

[31] Senthil Purushwalkam, Maximilian Nickel, Abhinav Gupta, and Marc'Aurelio Ranzato. Task-driven modular networks for zero-shot compositional learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3593–3602, 2019. 1, 4, 5, 6

[32] Frank Ruis, Gertjan Burghouts, and Doina Bucur. Independent prototype propagation for zero-shot compositionality. *Advances in Neural Information Processing Systems*, 34:10641–10653, 2021. 2

[33] 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. 1, 2

[34] Behjat Siddiquie, Rogerio S Feris, and Larry S Davis. Image ranking and retrieval based on multi-attribute queries. In *CVPR 2011*, pages 801–808. IEEE, 2011. 2

[35] Krishna Kumar Singh and Yong Jae Lee. End-to-end localization and ranking for relative attributes. In *European Conference on Computer Vision*, pages 753–769. Springer, 2016. 2

[36] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014. 5

[37] Chi Su, Shiliang Zhang, Junliang Xing, Wen Gao, and Qi Tian. Deep attributes driven multi-camera person re-identification. In *European conference on computer vision*, pages 475–491. Springer, 2016. 2

[38] Xin Wang, Fisher Yu, Ruth Wang, Trevor Darrell, and Joseph E Gonzalez. Tafe-net: Task-aware feature embeddings for low shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1831–1840, 2019. 5

[39] Guangyue Xu, Parisa Kordjamshidi, and Joyce Y Chai. Zero-shot compositional concept learning. *arXiv preprint arXiv:2107.05176*, 2021. 1, 2

[40] Ziwei Xu, Guangzhi Wang, Yongkang Wong, and Mohan S Kankanhalli. Relation-aware compositional zero-shot learning for attribute-object pair recognition. *IEEE Transactions on Multimedia*, 2021. 1

[41] Muli Yang, Cheng Deng, Junchi Yan, Xianglong Liu, and Dacheng Tao. Learning unseen concepts via hierarchical decomposition and composition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10248–10256, 2020. 1, 2

[42] Muli Yang, Chenghao Xu, Aming Wu, and Cheng Deng. A decomposable causal view of compositional zero-shot learning. *IEEE Transactions on Multimedia*, 2022. 2, 5, 6

[43] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 192–199, 2014. 4, 5

[44] Aron Yu and Kristen Grauman. Semantic jitter: Dense supervision for visual comparisons via synthetic images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5570–5579, 2017. 4, 5

[45] Tian Zhang, Kongming Liang, Ruoyi Du, Xian Sun, Zhanyu Ma, and Jun Guo. Learning invariant visual representations for compositional zero-shot learning. *arXiv preprint arXiv:2206.00415*, 2022. 2

[46] Xiao Zhang, Rui Zhao, Yu Qiao, Xiaogang Wang, and Hongsheng Li. Adacos: Adaptively scaling cosine logits for effectively learning deep face representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10823–10832, 2019. 4