Learning Attention as Disentangler for Compositional Zero-shot Learning

Shaozhe Hao Kai Han* Kwan-Yee K. Wong
The University of Hong Kong
{szhao, kykwong}@cs.hku.hk kaihanx@hku.hk

Abstract

*Compositional zero-shot learning (CZSL) aims at learning visual concepts (*i.e., *attributes and objects) from seen compositions and combining concept knowledge into unseen compositions. The key to CZSL is learning the disentanglement of the attribute-object composition. To this end, we propose to exploit cross-attentions as compositional disentanglers to learn disentangled concept embeddings. For example, if we want to recognize an unseen composition "yellow flower", we can learn the attribute concept "yellow" and object concept "flower" from different yellow objects and different flowers respectively. To further constrain the disentanglers to learn the concept of interest, we employ a regularization at the attention level. Specifically, we adapt the earth mover's distance (EMD) as a feature similarity metric in the cross-attention module. Moreover, benefiting from concept disentanglement, we improve the inference process and tune the prediction score by combining multiple concept probabilities. Comprehensive experiments on three CZSL benchmark datasets demonstrate that our method significantly outperforms previous works in both closed- and open-world settings, establishing a new state-of-the-art. Project page:* `https://haoosz.github.io/ade-czsl/`

1. Introduction

Suppose we have never seen white bears (*i.e.*, polar bears) before. Can we picture what it would look like? This is not difficult because we have seen many white animals in daily life (*e.g.*, white dogs and white rabbits) and different bears with various visual attributes in the zoo (*e.g.*, brown bears and black bears). Humans have no difficulty in disentangling "white" and "bear" from seen instances and combining them into the unseen composition. Inspired by this property of human intelligence, researchers attempt to make machines learn compositions of concepts as well. Compositional zero-shot learning (CZSL) is a specific problem studying visual compositionality, aiming to learn visual concepts from seen compositions of attributes and objects and generalize concept knowledge to unseen compositions.

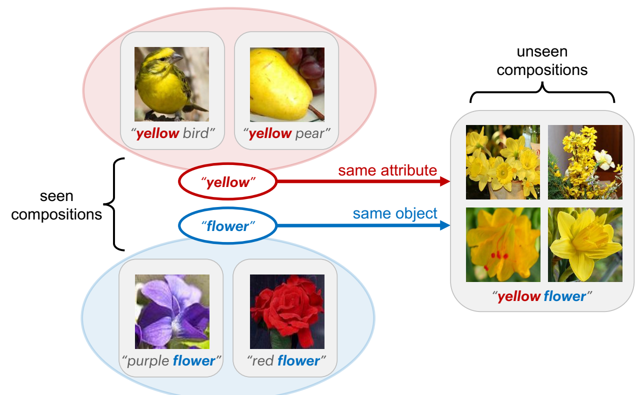

Figure 1. Motivation illustration. Given images from seen attribute-object compositions, human can disentangle the attribute "yellow" from "yellow bird" and "yellow pear", and the object "flower" from "purple flower" and "red flower". After learning visual properties of the concepts "yellow" and "flower", human can then recognize images from the unseen composition "yellow flower".

Learning attribute-object compositions demands prior knowledge about attributes and objects. However, visual concepts of attributes and objects never appear alone in a natural image. To learn exclusive concepts for compositionality learning, we need to disentangle the attribute concept and the object concept. As illustrated in Fig. 1, if we want to recognize the image of "yellow flower", it is necessary to learn the "yellow" concept and the "flower" concept, *i.e.*, disentangle visual concepts, from images of seen compositions. Previous works [22, 24, 25, 28–30, 36, 46] tackle CZSL by composing attribute and object word embeddings, and projecting word and visual embeddings to a joint space. They fail to disentangle visual concepts. Recently, some works [21,40,41,50] consider visual disentanglement but still have limitations despite their good performance. SCEN [21] learns concept-constant samples contrastively without constructing concept embedding prototypes to avoid learning irrelevant concepts shared by positive samples. IVR [50] disentangles visual features into ideal concept-invariant domains. This ideal domain generalization setting requires a small discrepancy of attribute and object sets and would degenerate on vaster and more

*Corresponding author

complex concepts. ProtoProp [40] and OADis [41] learn local attribute and object prototypes from spatial features on convolutional feature maps. However, spatial disentanglement is sometimes infeasible because attribute and object concepts are highly entangled in spatial features. Taking an image of "yellow flower" as an example, the spatial positions related to the attribute "yellow" and the object "flower" completely overlap, which hinders effective attribute-object disentanglement.

To overcome the above limitations, we propose a simple visual disentangling framework exploiting **A**ttention as **D**is**E**ntangler (ADE) on top of vision transformers [6]. We notice that vision transformers (ViT) have access to more sub-space global information across multi-head attentions than CNNs [38]. Therefore, with the expressivity of different subspace representations, token attentions of ViT may provide a more effective way for disentangling visual features, compared to using traditional spatial attentions across local positions on convolutional features [40, 41]. Specifically, it is difficult to disentangle the attribute-object composition "yellow flower" by spatial positions, but it is possible for ViT multi-head attentions to project attribute concept "yellow" and object concept "flower" onto different subspaces. Inspired by this property, we propose to learn cross-attention between two inputs that share the same concept, *e.g.*, "yellow bird" and "yellow pear" share the same attribute concept "yellow". In this way, we can derive attribute- and object-exclusive visual representations by cross-attention disentanglement. To ensure that the concept disentangler is exclusive to the specific concept, we also need to constrain the disentanglers to learn the concept of interest instead of the other concept. For example, given attribute-sharing images, the attribute attention should output similar attribute-exclusive features while the object attention should not. To achieve this goal, we apply a regularization term adapted from the earth mover's distance (EMD) [13] at the attention level. This regularization term forces cross-attention to learn the concept of interest by leveraging the feature similarity captured from all tokens. Mancini *et al.* [24] propose an open-world evaluation setting, which is neglected by most previous works. We consider both closed-world and the open-world settings in our experiments, demonstrating that our method is coherently efficient in both settings. The contributions of this paper are summarized below:

- We propose a new CZSL approach, named ADE, using cross-attentions to disentangle attribute- and object-exclusive features from paired concept-sharing inputs.
- We force attention disentanglers to learn the concept of interest with a regularization term adapted from EMD, ensuring valid attribute-object disentanglement.
- We comprehensively evaluate our method in both closed-world and open-world settings on three CZSL datasets, achieving consistent state-of-the-art.

2. Related work

Visual attribute has been widely studied to understand how visual properties can be learned from objects. The pioneering work by Ferrari and Zisserman [10] learned visual attributes using a probabilistic generative model. The successive work by Lampert *et al.* [20] used visual attributes to detect unseen objects with an attribute-based multi-label classification. Similarly, Patterson *et al.* [35] proposed Economic Labeling Algorithm (ELA) to discover multi-label attributes for objects. Different from multi-label classification, other works [4, 8, 15, 23] learned attribute-object relationship to generalize attribute feature across all object categories based on probabilistic models. Visual attributes also benefit downstream tasks, *e.g.*, object recognition [7, 16, 31], action recognition [1, 9, 26], image captioning [19, 33], and semi-supervised learning [42].

Compositional zero-shot learning (CZSL) is a special case of zero-shot learning (ZSL) [34, 39, 43, 47, 48], aims at recognizing unseen attribute-object compositions learning from seen compositions. Misra *et al.* [28] first termed and studied CZSL by projecting composed primitives and visual features to a joint embedding space. Nagarajan *et al.* [30] formulated attributes as matrix operators applied on object vectors. Purushwalkam *et al.* [36] introduced a task-driven modular architecture to learn unseen compositions by re-weighting a set of sub-tasks. Wei *et al.* [46] generated attribute-object compositions with GAN [11] to match visual features. Li *et al.* [22] proposed symmetry principle of attribute-object transformation under the supervision of group axioms. Naeem *et al.* [29] and Mancini *et al.* [25] used graph convolutional networks to extract attribute-object representations. Recently, some works shift their interest from word composing to visual disentanglement. Atzmon *et al.* [2] solved CZSL from a causal perspective to learn disentangled representations. Ruis *et al.* [40] proposed to learn prototypical representations of objects and attributes. Li *et al.* [21] disentangled visual features into a Siamese contrastive space and entangled them with a generative model. Saini *et al.* [41] extracted visual similarity from spatial features to disentangle attributes and objects. Zhang *et al.* [50] treated CZSL as a domain generalization task, learning attribute- and object-invariant domains. A more realistic open-world CZSL setting was studied in [17, 24, 25], which considered all possible compositions in testing. Very recently, an inspiring work by Nayak *et al.* [32] introduced compositional soft prompts to CLIP [37] to tackle CZSL problem.

Attention mechanism has been well studied by non-local neural networks [45] in computer vision and transformers [44] in machine translation. Dosovitskiy *et al.* [6] adapted transformer architecture to computer vision field, proving its comparable efficiency over traditional CNNs. Inspired by the multi-head self-attention implemented by transformers, our work exploits efficient attention as disentangler for CZSL.

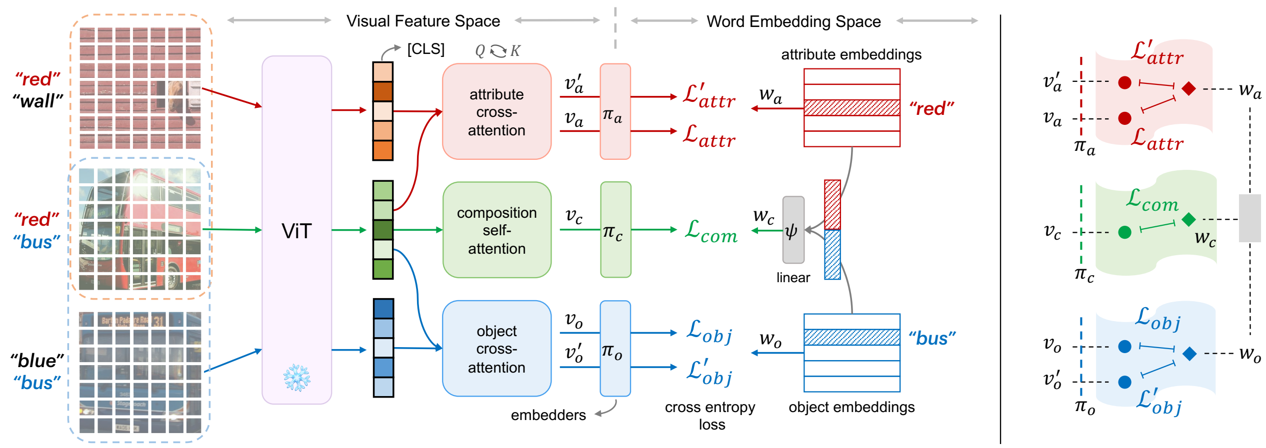

Figure 2. Method overview. Left (our framework ADE): Given one target image of "red bus", we sample two auxiliary images of the same attribute "red wall" and of the same object "blue bus". We feed the three images into a frozen ViT initialized with DINO [3]. We then input all encoded tokens (*i.e.*, `[CLS]` and patch tokens) to three attention modules: (1) attribute cross-attention taking paired attribute-sharing tokens as inputs; (2) object cross-attention taking paired object-sharing tokens as inputs; (3) composition self-attention taking tokens of single target image as input. We then project the `[CLS]` tokens of attention outputs with three MLP embedders π_a, π_c, and π_o. We finally compute cross entropy losses of embedded visual features with three learnable word embeddings: attribute embeddings, object embeddings, and their linear fused composition embeddings. Right: Illustration of five losses in attribute, object, and composition embedding spaces.

3. Preliminary

Compositional zero-shot learning aims at learning a model from limited compositions of attributes (*e.g.*, yellow) and objects (*e.g.*, flower) to recognize an image from novel compositions. Given a set of all possible attributes $\mathcal{A} = \{a_0, a_1, \cdots, a_n\}$ and a set of all possible objects $\mathcal{O} = \{o_0, o_1, \cdots, o_m\}$, a compositional class set $\mathcal{C} = \mathcal{A} \times \mathcal{O}$ includes all attribute-object compositions. We divide $\mathcal{C}$ into two disjoint sets, namely the set of seen classes $\mathcal{C}_s$ and the set of unseen classes $\mathcal{C}_u$, where $\mathcal{C}_s \cap \mathcal{C}_u = \varnothing$ and $\mathcal{C}_s \cup \mathcal{C}_u = \mathcal{C}$. Training images are only from classes in $\mathcal{C}_s$ and testing images are from classes in $\mathcal{C}_{test} = \mathcal{C}_s \cup \mathcal{C}'_u$. For closed-world evaluation, $\mathcal{C}'_u$ is a predefined subset of $\mathcal{C}_u$, *i.e.*, $\mathcal{C}'_u \subset \mathcal{C}_u$. For open-world evaluation, all possible compositions are considered for testing, *i.e.*, $\mathcal{C}'_u = \mathcal{C}_u$. CZSL task aims to learn a model $f : \mathcal{X} \to \mathcal{C}_{test}$ to predict labels in the testing composition set $c \in \mathcal{C}_{test}$ for input images $x \in \mathcal{X}$.

To learn a CZSL model, a common method is to minimize the cosine similarity score between the visually encoded feature and the word embedding of attribute-object compositions. The similarity score can be expressed as

$$s(x, (a, o)) = \frac{\phi(x)}{||\phi(x)||} \cdot \frac{\psi(a, o)}{||\psi(a, o)||} \tag{1}$$

where $\phi(\cdot)$ is the visual encoder, and $\psi(\cdot, \cdot)$ is the composition function. The CZSL model f can be formulated as

$$f(x; \theta_\phi, \theta_\psi) = \underset{a_i \in \mathcal{A}, o_j \in \mathcal{O}}{\arg\max}\ s(x, (a_i, o_j)) \tag{2}$$

The model f is optimized by learning the model parameters θ_ϕ and θ_ψ to match the input x with the most similar attribute-object composition. Visual encoder ϕ and composition function ψ, as two core components, take multiple forms in different methods. For visual encoder, works [40, 41, 50] manually design networks on top of a pretrained backbone (*e.g.*, ResNet18 [12]) to disentangle attribute and object embeddings for visual features. For composition functions, the object conditioned network [41], the graph embedding [29, 40], and the linear projector [24, 50] are commonly used. Unlike employing complex composition functions, we adopt the simple linear projector in our method.

4. Our approach

We aim to disentangle attribute and object features for visual representation by learning from image pairs sharing the same attributes or objects. For example, by learning the object "flower" from the "purple flower" and the "red flower" and learning the attribute "yellow" from the "yellow bird" and the "yellow pear", we can infer what a "yellow flower" looks like. To this end, we propose ADE to learn cross-attentions as concept disentanglers with the regularization of a novel token earth mover's distance. The whole framework of our method is shown in Fig. 2.

4.1. Disentanglement with cross-attention

Multi-head self-attention in transformers [44] is powerful for extracting token connections. Attention maps each query token, and key-value token pairs to an output token. Every output token is a weighted sum of the values, where the weights is the similarity of the corresponding query token and the key tokens. The attention is formulated as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{3}$$

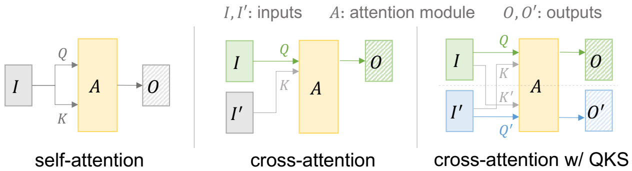

Figure 3. Illustration of three types of attention.

where three inputs are query tokens (Q), key tokens (K), and value tokens (V) and the scaling factor d_k is the dimension of the query and key. Self-attention uses the same input for Q, K, V and effectively captures relationships among [CLS] and patch tokens of a single input image. Normally, the [CLS] token is exclusively used as a global representation of the input image for downstream tasks, but exclusive [CLS] token highly entangles attribute and object concepts.

For CZSL task, we want to disentangle the exclusive concept by exploiting the semantic similarity between tokens from different concept-sharing features. For this purpose, we introduce a *cross*-attention mechanism to extract attribute-exclusive or object-exclusive features for paired images sharing the same attribute or object concept. The cross-attention has the same structure as the self-attention, but works on different inputs. Self-attention uses the same input I for Q, K, V (*i.e.*, Q=K=V=I) while cross-attention uses one of the paired inputs I for the query and the other I' for the key and the value (*i.e.*, Q=I, K=V=I'). In the cross-attention, the query input and the key input play different roles. The cross-attention output is the weighted sum of the value (also the key) input based on the similarity between the query and key inputs. For a pair of input images, we swap the query and key inputs in computing cross-attention to derive two cross-attended outputs for the respective inputs. Cross-attention with query-key swapping (QKS) is illustrated in Fig. 3.

The cross-attention outputs paired features of the exclusive concept with paired concept-sharing inputs. Taking the object-sharing pair as an example, object cross-attention leverages the shared object features of the key input to enhance the query input, thus making the outputs exclusive to object feature space, *i.e.*, object-exclusive features v_o and v'_o. The same applies to attribute-exclusive features v_a and v'_a extracted from paired attribute-sharing images. We also use a standard self-attention to extract composition feature v_c.

To train cross-attentions, we adopt common cross entropy loss after MLP embedders (π_a, π_c, π_o) for attribute-exclusive features (v_a, v'_a), object-exclusive features (v_o, v'_o), and composition feature (v_c). We denote any of these visual features as v and the corresponding embedder as π. We denote the word embedding as w_z, where the concept $z \in \mathcal{Z}$ can be any of the attribute $a \in \mathcal{A}$, the object $o \in \mathcal{O}$, or the composition $c \in \mathcal{C}$. The class probability of L2-normalized $\pi(v)$ and w_z can be computed as:

$$p_\pi(z \mid v) = \frac{\exp\left(\pi(v) \cdot w_z/\tau\right)}{\sum_{\hat{z}\in\mathcal{Z}} \exp\left(\pi(v) \cdot w_{\hat{z}}/\tau\right)}, z \in \mathcal{Z} \tag{4}$$

where τ is the temperature. We then compute the cross entropy to classify visual features:

$$H_\pi(v, z) = -\log p_\pi(z \mid v), z \in \mathcal{Z} \tag{5}$$

Considering different classification objectives of attribute, object, and composition, the general Eq. (5) can be applied to five visual features, *i.e.*, v_a, v'_a, v_o, v'_o, v_c. Therefore, the training objective is composed of five cross entropy losses:

$$\begin{aligned}\mathcal{L}_{ce} = &\underbrace{H_{\pi_a}(v_a, a)}_{\mathcal{L}_{attr}} + \underbrace{H_{\pi_a}(v'_a, a)}_{\mathcal{L}'_{attr}} + \underbrace{H_{\pi_c}(v_c, c)}_{\mathcal{L}_{com}} + \\ &\underbrace{H_{\pi_o}(v_o, o)}_{\mathcal{L}_{obj}} + \underbrace{H_{\pi_o}(v'_o, o)}_{\mathcal{L}'_{obj}}\end{aligned} \tag{6}$$

where the attribute label $a \in \mathcal{A}$, the object label $o \in \mathcal{O}$, and the composition label $c = (a, o) \in \mathcal{C}_{test}$.

4.2. Disentangling constraint at the attention level

So far, we have enabled cross-attentions to disentangle concepts, but we want to further ensure that the attribute and the object disentanglers learn the corresponding attribute and object concepts instead of the opposite concepts. To this end, we introduce an attention-level earth mover's distance (EMD) to constrain disentanglers to learn the concept of interest. The EMD is formulated as an optimal transportation problem in [13]. Suppose we have supplies of n_s sources $\mathcal{S} = \{s_i\}_{i=1}^{n_s}$ and demands of n_d destinations $\mathcal{D} = \{d_j\}_{j=1}^{n_d}$. Given the moving cost from the i-th source to the j-th destination c_{ij}, an optimal transportation problem aims to find the minimal-cost flow f_{ij} from sources to destinations:

$$\underset{f_{ij}}{\text{minimize}} \quad \sum_{i=1}^{n_s} \sum_{j=1}^{n_d} c_{ij} f_{ij} \tag{7}$$

$$\text{subject to} \quad f_{ij} \geqslant 0,\ i = 1, ..., n_s,\ j = 1, ..., n_d \tag{8}$$

$$\sum_{j=1}^{n_d} f_{ij} = s_i,\ i = 1, ..., n_s \tag{9}$$

$$\sum_{i=1}^{n_s} f_{ij} = d_j,\ j = 1, ..., n_d \tag{10}$$

where the optimal flow $\tilde{f}_{ij}$ is computed by the moving cost c_{ij}, the supplies s_i, and the demands d_j. The EMD can be further formulated as:

$$\text{EMD}(c_{ij}, s_i, d_j) = (1 - c_{ij})\tilde{f}_{ij}. \tag{11}$$

When the EMD is greater, the distributions $\mathcal{S}$ and $\mathcal{D}$ are closer. In Fig. 4, we show how to use the EMD as a feature similarity metric at the attention level in our cross-attention module. We can derive a cross-attention map from Eq. (3) and the counterpart cross-attention map after query-key swapping. We use the logits of [CLS]-to-patch attentions as supplies s_i and demands d_j, which represent how similar the global feature of one image is to the local patches of the

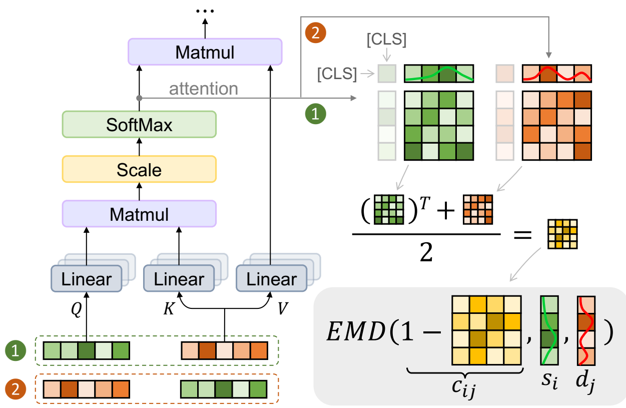

Figure 4. Illustration of the adapted EMD at the attention level. Branches ❶ and ❷ respectively denote using one of the paired inputs as the query Q. We compute the EMD with two attention maps from ❶ and ❷ branches. We use the `[CLS]`-to-patch logits as the supplies s_i and the demands d_j, and one minus the mean matrix of patch-to-patch logits as the moving cost c_{ij}. In this way, we adapt the EMD to our cross-attention module.

other image. We use one minus the average map of patch-to-patch attentions as the moving cost c_{ij}. In this way, we can compute an adapted EMD λ for the cross-attention:

$$\lambda_q^p = \text{EMD}(c_{ij}, s_i, d_j; p, q), p, q \in \{a, o\} \tag{12}$$

where p represents the type of the cross-attention, *i.e.*, attribute cross-attention when $p = a$ and object cross-attention when $p = o$, and q represents the type of the inputs, *i.e.*, attribute-sharing inputs when $q = a$ and object-sharing inputs when $q = o$.

For attribute cross-attention, attribute-sharing inputs should have large EMD λ_a^a while object-sharing inputs should have small EMD λ_o^a. It is opposite for object cross-attention, *i.e.*, small λ_a^o of attribute-sharing inputs and large λ_o^o of object-sharing inputs. Thus, we formulate the regularization term as:

$$\mathcal{L}_{reg} = \lambda_o^a + \lambda_a^o - \lambda_a^a - \lambda_o^o \tag{13}$$

4.3. Training and inference

At the training phase, we formulate our final loss as a cross entropy with regularization for all inputs:

$$\mathcal{L} = \mathcal{L}_{ce} + \mathcal{L}_{reg} \tag{14}$$

At inference phase, we feed the same test input into three attention branches in a self-attention manner, deriving three class probabilities $p(c)$, $p(a)$, and $p(o)$[1], where $c = (a, o)$. Unlike most methods modeling single $p(c)$ for prediction, we compute our prediction score by synthesizing attribute, object and composition probabilities:

$$\hat{c} = \underset{c \in \mathcal{C}_{test}}{\arg\max}\ p(c) + \beta \cdot p(a) \cdot p(o) \tag{15}$$

[1]Abbreviations for $p_{\pi_c}(c \mid v_c)$, $p_{\pi_a}(a \mid v_a)$, and $p_{\pi_o}(o \mid v_o)$, where v_c, v_a, and v_o are outputs from three self-attentions.

We first fix $\beta = 1.0$ during training and then validate $\beta = 0.0, 0.1, \cdots, 1.0$ to choose the best β on the validation set. Finally, we compute the composition prediction as Eq. (15) with the chosen β, making the best prediction trade-off between generalized composition and independent concept.

5. Experiments

5.1. Experimental details

Datasets We use three benchmark datasets in CZSL problem, namely Clothing16K [50], UT-Zappos50K [49], and C-GQA [29]. Clothing16K [50] contains different types of clothing (*e.g.*, shirt, pants) with color attributes (*e.g.*, white, black). UT-Zappos50K [49] is a fine-grained dataset consisting of different kinds of shoes (*e.g.*, sneakers, sandals) with texture attributes (*e.g.*, leather, canvas). C-GQA [29] is a split built on top of Stanford GQA dataset [14], composed of extensive common attribute concepts (*e.g.*, old, wet) and object concepts (*e.g.*, dog, bus) in real life. We follow the common data splits of these three datasets (see Table 1).

	Composition			Train		Val		Test	
Datasets	$\lvert\mathcal{A}\rvert$	$\lvert\mathcal{O}\rvert$	$\lvert\mathcal{A}\rvert \times \lvert\mathcal{O}\rvert$	$\lvert\mathcal{C}_s\rvert$	$\lvert\mathcal{X}\rvert$	$\lvert\mathcal{C}_s\rvert$ / $\lvert\mathcal{C}_u\rvert$	$\lvert\mathcal{X}\rvert$	$\lvert\mathcal{C}_s\rvert$ / $\lvert\mathcal{C}_u\rvert$	$\lvert\mathcal{X}\rvert$
Clothing16K [50]	9	8	72	18	7242	10 / 10	5515	9 / 8	3413
UT-Zappos50K [49]	16	12	192	83	22998	15 / 15	3214	18 / 18	2914
C-GQA [29]	413	674	278362	5592	26920	1252 / 1040	7280	888 / 923	5098

Table 1. Summary of data split statistics.

Open-world setting In addition to the standard closed-world setting, we also evaluate our model on the open-world setting [24], which is neglected by most previous works. The open-world setting considers all possible compositions, which requires a much larger testing space than the closed-world setting during inference. Taking UT-Zappos50K as an example (see Table 1), the closed world only considers 36 compositions in the testing set while the open world considers total 192 compositions, in which ∼81% are ignored under the standard closed-world setting.

Evaluation metrics Since CZSL models have an inherent bias for seen compositions, we follow the generalized CZSL evaluation protocol [36]. To overcome the negative bias for seen compositions, we apply different calibration terms to unseen compositions and compute the corresponding top-1 accuracy of seen and unseen compositions, where a larger bias makes higher unseen accuracy and lower seen accuracy, and vice versa. We treat seen accuracy as x-axis and unseen accuracy as y-axis to derive an unseen-seen accuracy curve. We can then compute the area under curve (AUC), the best harmonic mean, the best seen accuracy, and the best unseen accuracy from the curve. In our experiments, we report these four metrics for evaluation, among which AUC is the most representative and stable metric for measuring CZSL model performance. Note that the attribute accuracy or the object accuracy alone does not reflect CZSL performance, because the individual accuracy on attribute or object does not necessarily decide the accuracy of their composition.

Closed-world	Clothing16K						UT-Zappos50K						C-GQA					
Models	AUC	HM	Seen	Unseen	Attr	Obj	AUC	HM	Seen	Unseen	Attr	Obj	AUC	HM	Seen	Unseen	Attr	Obj
SymNet [22]	78.8	79.3	98.0	85.1	75.6	84.1	32.6	45.6	60.6	68.6	48.2	77.0	3.1	13.5	30.9	13.3	11.4	34.6
CompCos [24]	90.3	87.2	98.5	96.8	**90.2**	91.8	31.8	48.1	58.8	63.8	45.5	72.4	2.9	12.8	30.7	12.2	10.4	33.9
GraphEmb [29]	89.2	84.2	98.0	97.4	90.0	93.1	34.5	48.5	61.6	**70.0**	**50.8**	**77.1**	3.8	15.0	32.3	14.9	13.8	33.2
Co-CGE [25]	88.3	87.9	98.5	94.7	87.4	91.4	30.8	44.6	60.9	62.6	46.0	73.5	3.6	14.7	31.6	14.3	12.6	34.6
SCEN [21]	78.8	78.5	98.0	89.6	81.2	85.4	30.9	46.7	**65.7**	62.9	44.0	74.4	3.5	14.6	31.7	13.4	10.7	31.4
IVR [50]	90.6	86.6	**99.0**	97.0	89.3	**93.6**	34.3	49.2	61.5	68.1	48.4	74.6	2.2	10.9	27.3	10.0	10.3	**37.5**
OADis [41]	88.4	86.1	97.7	94.2	84.9	93.1	32.6	46.9	60.7	68.8	49.3	76.9	3.8	14.7	33.4	14.3	8.9	36.3
ADE (ours)	**92.4**	**88.7**	98.2	**97.7**	**90.2**	**93.6**	**35.1**	**51.1**	63.0	64.3	46.3	74.0	**5.2**	**18.0**	**35.0**	**17.7**	**16.8**	32.3

Table 2. Closed-world results on three datasets. We report the area under curve (AUC), the best harmonic mean (HM), the best seen accuracy (Seen), the best unseen accuracy (Unseen), the attribute accuracy (Attr), and the object accuracy (Obj) of the unseen-seen accuracy curve under the closed world-setting. AUC is the core CZSL metric. All models use the same DINO ViT-B-16 backbone.

Open-world	Clothing16K						UT-Zappos50K						C-GQA					
Models	AUC	HM	Seen	Unseen	Attr	Obj	AUC	HM	Seen	Unseen	Attr	Obj	AUC	HM	Seen	Unseen	Attr	Obj
SymNet [22]	57.4	68.3	98.2	60.7	57.6	81.2	25.0	40.6	60.4	51.0	38.2	**75.0**	0.77	4.9	30.1	3.2	18.4	37.5
CompCos [24]	64.1	70.8	98.2	69.8	71.7	83.7	20.7	36.0	58.1	46.0	36.4	71.1	0.72	4.3	32.8	2.8	15.1	37.8
GraphEmb [29]	62.0	68.3	98.5	69.7	71.8	82.4	23.5	40.0	60.6	47.0	37.1	69.3	0.81	4.8	32.7	3.2	17.2	36.7
Co-CGE [25]	59.3	69.2	98.7	63.8	68.5	76.2	22.0	40.3	57.7	43.4	33.9	67.2	0.48	3.3	31.1	2.1	15.5	35.7
SCEN [21]	53.7	61.5	96.7	62.3	63.6	79.1	22.5	38.0	**64.8**	47.5	34.9	73.3	0.34	2.5	29.5	1.5	14.8	32.3
IVR [50]	63.6	72.0	98.7	69.0	70.3	84.8	25.3	42.3	60.7	50.0	38.4	71.4	0.94	5.7	30.6	4.0	16.9	36.5
OADis [41]	53.4	63.2	98.0	58.6	57.3	**85.4**	25.3	41.6	58.7	**53.9**	**40.3**	74.7	0.71	4.2	33.0	2.6	14.6	**39.7**
ADE (ours)	**68.0**	**74.2**	**99.0**	**73.1**	**75.0**	84.5	**27.1**	**44.8**	62.4	50.7	39.9	71.4	**1.42**	**7.6**	**35.1**	**4.8**	**22.4**	35.6

Table 3. Open-world results on three datasets. Different from Table 2, open-world setting considers all possible compositions in testing.

Implementation details We use a frozen ViT-B-16 [6] backbone pretrained with DINO [3] on ImageNet [5] in a self-supervised manner as our visual feature extractor. The ViT-B-16 outputs contain 197 tokens (1 `[CLS]` and 196 patch tokens) of 768 dimensions. For three attention disentangler modules, we implement one-layer multi-head attention framework following [44], changing the single input to paired inputs for cross-attentions. The embedders π_a, π_c, π_o are the two-layer MLPs following the previous works [24, 50], projecting the 768-dimension visual features to 300-dimension word embedding space. The word embedding prototypes are initialized with word2vec [27] for all datasets and learnable during training. The composition function ψ is one linear layer. We train our model using Adam optimizer [18] with a learning rate of 5×10^{-6} for UT-Zappos50K and Clothing16K, and 5×10^{-5} for C-GQA. All models are trained with 128 batch size for 300 epochs.

5.2. Comparison

To ensure a fair comparison and demonstrate that our improvement over baseline models is not merely by ViT, we adopt ViT backbone to state-of-the-art CZSL models and *re-train* all models. We compare our method with them: (1) OADis [41] disentangles attribute and object features from spatial convolutional maps; (2) SymNet [22] introduces the symmetry principle of attribute-object transformation and group theory as training objectives; (3) CompCos [24] extends CZSL to an open-world setting considering all possible compositions during inference, proposing a feasibility score based on data statistics to remove unfeasible compositions; (4) GraphEmb [29] and Co-CGE [25] propose to use graph convolutional networks (GCN) to represent attribute-object relationships and compositions; (5) SCEN [21] projects visual features to a Siamese contrastive space to capture concept prototypes, and introduces complex state transition module to produce virtual compositions; (6) IVR [50] proposes to disentangle visual features into concept-invariant domains from a perspective of domain generalization, by masking specific channels of visual features.

Closed-world evaluation In Table 2, we compare our ADE model with the state-of-the-art methods. ADE consistently outperforms others by a significant margin. ADE increases the core metric AUC by 1.8 on Clothing16K, 0.6 on UT-Zappos50K, and 1.4 on C-GQA (∼37% relatively). Similarly, ADE increases the best harmonic mean (HM) by 0.8% on Clothing16K, 1.9% on UT-Zappos50K, and 3.0% on C-GQA. We notice that SymNet [22] and SCEN [21] perform badly on Clothing16K. The reason might be that not learning concept prototypes harms the word embedding expressivity on small-scale concepts. We also notice that IVR [50] performs very well on curated datasets Clothing16K and UT-Zappos50K but badly on larger-scale real-world dataset C-GQA. We hypothesize ideal concept-invariant domains might be difficult to learn from natural images and large-scale concepts of C-GQA. In contrast, our ADE model achieves state-of-the-art performance on all datasets.

Open-world evaluation In Table 3, we consider the open-world setting to compare our ADE with other methods. Likewise, ADE also performs the best among all methods under open-world setting. ADE increases AUC by 3.9 on Cloth-

	CA	AA	OA	Reg	AUC	HM	Seen	Unseen
(0)	✗	✗	✗	✗	23.8	41.1	59.0	48.9
(1)	self	✗	✗	✗	25.3	42.3	61.1	49.9
(2)	self	self	self	✗	26.7	44.6	61.9	49.8
(3)	self	cross	cross	✗	26.9	44.5	**63.4**	48.7
(4)	self	cross	cross	✓	**27.1**	**44.8**	62.4	**50.7**

Table 4. Ablate the components in ADE on open-world UT-Zappos50K. CA, AA, and OA denote composition, attribute, and object attention. Reg denotes the regularization term. We test self- or cross-attention for AA and OA.

		C-GQA				Clothing16K			
	Inference formulation	AUC	HM	Seen	Unseen	AUC	HM	Seen	Unseen
(0)	$p(c)$	4.6	16.8	**35.1**	16.0	**92.4**	**88.8**	**98.2**	**97.7**
(1)	$p(a) \cdot p(o)$	4.0	15.8	31.4	15.1	57.3	66.3	96.7	63.1
(2)	$p(c) + p(a) \cdot p(o)$	**5.2**	**18.0**	35.0	**17.7**	90.4	85.9	98.2	97.0
(3)	$p(c) + \beta \cdot p(a) \cdot p(o)$	**5.2**	**18.0**	35.0	**17.7**	**92.4**	88.7	**98.2**	**97.7**

Table 5. Results on closed-world Clothing16K and C-GQA using different inference formulations. Rows (0)-(2) respectively represents the cases when $\beta = 0.0$, $\beta = +\infty$, and $\beta = 1.0$. Row (3) is our inference formulation, which applies an β optimized on the validation set.

ing16K, 1.8 on UT-Zappos50K, and 0.48 on C-GQA ($\sim$51% relatively). ADE also increases the best harmonic mean (HM) by 2.2% on Clothing16K, 2.5% on UT-Zappos50K, and 1.9% on C-GQA ($\sim$33% relatively). From the above results, ADE surpasses others by a larger margin on open-world AUC and HM than closed-world ones, indicating ADE maintains utmost efficiency when turning from the closed world to the open world. It is worth mentioning that ADE does not apply any special operations (*e.g.*, feasibility masking [24]) for the open world and deals with the two settings in exactly the same way. IVR [50] keeps its performance to a great extent but still lags behind our method significantly.

5.3. Ablation study

Backbone: ResNet *vs* ViT Our work leverages ViT as the default backbone to excavate more high-level sub-space information, while ResNet18 is the most common backbone in previous works. In Table 6, we compare our ADE to OADis [41] with both backbones. Our ADE performs similarly to OADis with ResNet18, but outperforms it significantly with ViT. Additionally, we present an ablation study on different components of our method with the ResNet18 backbone in the Appendix. These experiments indicate that our model benefits from ViT and all components of our method are effective regardless of the backbone.

Closed-world		Clothing16K		UT-Zappos50K		C-GQA	
Backbone	Models	AUC	HM	AUC	HM	AUC	HM
ResNet18	OADis [41]	85.5	84.7	**30.0**	44.4	**3.1**	13.6
	ADE (ours)	**87.2**	**85.1**	29.5	**47.0**	**3.1**	**13.7**
ViT-B-16	OADis [41]	88.4	86.1	32.6	46.9	3.8	14.7
	ADE (ours)	**92.4**	**88.7**	**35.1**	**51.1**	**5.2**	**18.0**

Table 6. Compare ADE and OADis [41] with ResNet18 and ViT.

Different parts of ADE We evaluate the effectiveness of attention disentanglers (composition, attribute, and object attention) and the regularization term in our model. We report the ablation study results on the open-world UT-Zappos50K in Table 4. Rows (0)-(2) show attention disentanglers can significantly improve the performance. Rows (2)-(3) show that cross-attention learns disentangled concepts better than self-attention for AA and OA. Rows (3)-(4) show the regularization term can further benefit the visual disentanglement, improving the unseen accuracy and overall AUC.

Inference formulation We also investigate the effect of our inference formulation $p(c) + \beta \cdot p(a) \cdot p(o)$ in Table 5. We report the results with extreme values of β, *i.e.*, $\beta = 0.0$ and $\beta = 1.0$. Note that $\beta = 0.0$ means only using composition probability for prediction. In addition, we also test the performance only using the product of attribute and object probabilities $p(a) \cdot p(o)$. We can observe that the best fixed β value is unfixed among datasets. For example, $\beta = 1.0$ gives the highest AUC for C-GQA in row (2) while $\beta = 0.0$ for Clothing16K in row (0). In contrast, our validated β consistently gives the best inference results for both datasets. Another observation on C-GQA is that $p(a) \cdot p(o)$ alone is not a good prediction, but adding it to $p(c)$ can increase the unseen accuracy. This indicates that the disentangled attribute prediction $p(a)$ and object prediction $p(o)$ indeed enhance the unseen generalization for CZSL problem.

Effect of regularization term We propose an EMD-adapted regularization term at the attention level to force attentions to disentangle the concept of interest. We also investigate the effect of applying the regularization term at the feature level. Specifically, we compare our EMD-based distance to the cosine and euclidean feature distances. The results on open-world UT-Zappos50K are shown in Table 7. Our EMD-based regularization outperforms other distance forms, because our attention-level EMD distance considers token-wise similarity capturing the specific concept-related attention responses.

Reg	AUC	HM	Seen	Unseen
Cosine	26.8	44.7	**63.0**	48.6
Euclidean	26.2	44.3	62.6	47.5
Ours (EMD)	**27.1**	**44.8**	62.4	**50.7**

Table 7. Comparison of different regularization terms on open-world UT-Zappos50K.

5.4. Qualitative analysis

Visual disentanglement in feature space is hard to visualize [41]. Inspired by previous work attempts [22, 30, 41, 50], we conduct qualitative analysis of image and text retrieval to show how our ADE model correlates the visual image and the concept composition. In addition, to further validate ADE is efficient to disentangle visual concepts, we conduct unseen-to-seen image retrieval based on their visual concept features extracted by attribute and object attentions.

(a) Top-5 text-to-image retrieval. (b) Top-5 image-to-text retrieval. (c) Top-5 visual concept retrieval.

Figure 5. Qualitative analysis. (a) In the last row of "suede sandals", the wrong image (red box) is "fake leather sandals". (b) Each image has the ground-truth label (black text) and 5 retrieval results (colored text), in which the green text is the correct prediction. (c) We retrieve images sharing the same visual concepts by their visual concept features for unseen images of "yellow skirt" and "pink pants".

Image and text retrieval We first consider text-to-image retrieval. Given a text composition, *e.g.*, "leather heels", we embed it and retrieve the top-5 closest visual features based on the feature distance. We display four text compositions of the different objects sharing the same attributes and vice versa in Fig. 5a. We can observe that the retrieved images are correct in most of the cases. One exception is when retrieving "suede sandals", the third closest image is "fake leather sandals". Although "suede sandals" and "fake leather sandals" are not the same composition, they are quite visually similar. We then consider image-to-text retrieval, shown in Fig. 5b. Given an image, *e.g.*, the image of a "brown zebra", we extract its visual feature and retrieve the top-5 closest text composition embeddings. It is difficult to retrieve the ground-truth label in the top-1 closest text composition, but all top-5 results are all semantically related to the image. We take the image of "blond person" (row 3, col 2) as an example. Although the text composition "blond person" is not retrieved in the top-5 matches, the retrieved results "white shirts", "white outfit", "white shorts", "white pants", and "young girl" are all reasonable and actually present in the image. Image and text retrieval experiments validate that our ADE efficiently projects visual features and word embeddings into a uniform space.

Visual concept retrieval Because the attribute and the object are visually coupled in an image to a high degree of entanglement, it is challenging to visualize the disentanglement in feature space [41]. Saini *et al.* [41] retrieve single attribute or object text from test images. However, this process is the same as multi-label classification and insufficient to validate that disentangled visual concepts are learned from images. Based on the disentanglement ability of ADE, we construct a visual concept retrieval experiment to investigate the distances between visual concept features, *i.e.*, the embedded attribute feature $\pi_a(v_a)$ and the embedded object feature $\pi_o(v_o)$, extracted from different images. Prior to our work, no existing models can do so, because none of them extracts concept-exclusive features like ADE. The results are shown in Fig. 5c. We first extract attribute features and object features from all seen images. Given an unseen image, we retrieve the top-5 closest images by measuring the feature distance between the attribute feature of the given image and that of all seen images, and the same goes for the object feature. For the image of "yellow skirt", all retrieval results for the visual concept "yellow" are all "yellow `[OBJ]`", and all retrieval results for "skirt" are "`[ATTR]` skirt". For the "pink pants" image, our model also perfectly retrieves the visual concepts, *i.e.*, the attribute "pink" and the object "pants". Our experimental results demonstrate that our ADE model is effective to disentangle visual concepts from seen compositions and combine learned concept knowledge into unseen compositions.

6. Conclusion

In this paper, we tackle the problem of compositional zero-shot learning (CZSL). To disentangle visual concepts from the attribute-object composition, we propose ADE adopting cross-attentions to learn the individual concept from paired concept-sharing images. To constrain the disentanglers to learn the concept of interest, we employ a regularization term adapted from the earth mover's distance (EMD), which is used as a feature similarity metric in the cross-attention module. Moreover, to exploit the attribute and object prediction ability of ADE, we improve the inference process by combining attribute, object, and composition probabilities into the final prediction score. We empirically demonstrate ADE outperforms the current state-of-the-art methods under both closed- and open-world settings. We also conduct a comprehensive qualitative analysis to validate the disentanglement ability of attention disentanglers in ADE.

Limitations Like existing CZSL methods, it is time- and computation-consuming to derive all composition embeddings when the numbers of attributes and objects are large. Moreover, it remains an open challenge to exploit concepts based on the actual semantics rather than solely on text; for instance, the "open" attribute in "open curtain" and "open computer" has completely different meanings.

Acknowledgements This work is partially supported by Hong Kong Research Grant Council - Early Career Scheme (Grant No. 27208022) and HKU Seed Fund for Basic Research.

References

[1] Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Simon Lacoste-Julien. Joint discovery of object states and manipulation actions. In *ICCV*, 2017. 2

[2] Yuval Atzmon, Felix Kreuk, Uri Shalit, and Gal Chechik. A causal view of compositional zero-shot recognition. In *NeurIPS*, 2020. 2

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 3, 6

[4] Chao-Yeh Chen and Kristen Grauman. Inferring analogous attributes. In *CVPR*, 2014. 2

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 6

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2, 6

[7] Ali Farhadi, Ian Endres, and Derek Hoiem. Attribute-centric recognition for cross-category generalization. In *CVPR*, 2010. 2

[8] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *CVPR*, 2009. 2

[9] Alireza Fathi and James M Rehg. Modeling actions through state changes. In *CVPR*, 2013. 2

[10] Vittorio Ferrari and Andrew Zisserman. Learning visual attributes. *NeurIPS*, 2007. 2

[11] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 2

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[13] Frank L. Hitchcock. The distribution of a product from several sources to numerous localities. *Journal of Mathematics and Physics*, 1941. 2, 4

[14] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 5

[15] Sung Ju Hwang, Fei Sha, and Kristen Grauman. Sharing features between objects and their attributes. In *CVPR*, 2011. 2

[16] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *CVPR*, 2015. 2

[17] Shyamgopal Karthik, Massimiliano Mancini, and Zeynep Akata. Kg-sp: Knowledge guided simple primitives for open world compositional zero-shot learning. In *CVPR*, 2022. 2

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[19] 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 TPAMI*, 2013. 2

[20] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 2

[21] Xiangyu Li, Xu Yang, Kun Wei, Cheng Deng, and Muli Yang. Siamese contrastive embedding network for compositional zero-shot learning. In *CVPR*, 2022. 1, 2, 6

[22] Yong-Lu Li, Yue Xu, Xiaohan Mao, and Cewu Lu. Symmetry and group in attribute-object compositions. In *CVPR*, 2020. 1, 2, 6, 7

[23] Dhruv Mahajan, Sundararajan Sellamanickam, and Vinod Nair. A joint learning framework for attribute models and object descriptions. In *ICCV*, 2011. 2

[24] M Mancini, MF Naeem, Y Xian, and Zeynep Akata. Open world compositional zero-shot learning. In *CVPR*, 2021. 1, 2, 3, 5, 6, 7

[25] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Learning graph embeddings for open world compositional zero-shot learning. *IEEE TPAMI*, 2022. 1, 2, 6

[26] Tomas McCandless and Kristen Grauman. Object-centric spatio-temporal pyramids for egocentric activity recognition. In *BMVC*, 2013. 2

[27] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NeurIPS*, 2013. 6

[28] Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. In *CVPR*, 2017. 1, 2

[29] MF Naeem, Y Xian, F Tombari, and Zeynep Akata. Learning graph embeddings for compositional zero-shot learning. In *CVPR*, 2021. 1, 2, 3, 5, 6

[30] Tushar Nagarajan and Kristen Grauman. Attributes as operators: factorizing unseen attribute-object compositions. In *ECCV*, 2018. 1, 2, 7

[31] Zhixiong Nan, Yang Liu, Nanning Zheng, and Song-Chun Zhu. Recognizing unseen attribute-object pair with generative model. In *AAAI*, 2019. 2

[32] 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

[33] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. In *NeurIPS*, 2011. 2

[34] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. Zero-shot learning with semantic output codes. In *NeurIPS*, 2009. 2

[35] Genevieve Patterson and James Hays. Coco attributes: Attributes for people, animals, and objects. In *ECCV*, 2016. 2

[36] Senthil Purushwalkam, Maximilian Nickel, Abhinav Gupta, and Marc'Aurelio Ranzato. Task-driven modular networks for zero-shot compositional learning. In *ICCV*, 2019. 1, 2, 5

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2

[38] Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *NeurIPS*, 2021. 2

[39] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *ICML*, 2015. 2

[40] Frank Ruis, Gertjan Burghouts, and Doina Bucur. Independent prototype propagation for zero-shot compositionality. In *NeurIPS*, 2021. 1, 2, 3

[41] Nirat Saini, Khoi Pham, and Abhinav Shrivastava. Disentangling visual embeddings for attributes and objects. In *CVPR*, 2022. 1, 2, 3, 6, 7, 8

[42] Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Constrained semi-supervised learning using attributes and comparative attributes. In *ECCV*, 2012. 2

[43] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In *NeurIPS*, 2013. 2

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2, 3, 6

[45] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 2

[46] Kun Wei, Muli Yang, Hao Wang, Cheng Deng, and Xianglong Liu. Adversarial fine-grained composition learning for unseen attribute-object recognition. In *ICCV*, 2019. 1, 2

[47] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *CVPR*, 2018. 2

[48] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *CVPR*, 2017. 2

[49] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *CVPR*, 2014. 5

[50] Tian Zhang, Kongming Liang, Ruoyi Du, Xian Sun, Zhanyu Ma, and Jun Guo. Learning invariant visual representations for compositional zero-shot learning. In *ECCV*, 2022. 1, 2, 3, 5, 6, 7