Learning Unseen Concepts via Hierarchical Decomposition and Composition

Muli Yang[1], Cheng Deng[1]*, Junchi Yan[2], Xianglong Liu[3], and Dacheng Tao[4]
[1]School of Electronic Engineering, Xidian University, Xi'an 710071, China
[2]Department of CSE, and MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University, China
[3]State Key Lab of Software Development Environment, Beihang University, Beijing 100191, China
[4]UBTECH Sydney AI Centre, School of CS, Faculty of Engineering, The University of Sydney, Australia
mlyang@stu.xidian.edu.cn, chdeng@mail.xidian.edu.cn, yanjunchi@sjtu.edu.cn,
xlliu@nlsde.buaa.edu.cn, dacheng.tao@sydney.edu.au

Abstract

Composing and recognizing new concepts from known sub-concepts has been a fundamental and challenging vision task, mainly due to 1) the diversity of sub-concepts and 2) the intricate contextuality between sub-concepts and their corresponding visual features. However, most of the current methods simply treat the contextuality as rigid semantic relationships and fail to capture fine-grained contextual correlations. We propose to learn unseen concepts in a hierarchical decomposition-and-composition manner. Considering the diversity of sub-concepts, our method decomposes each seen image into visual elements according to its labels, and learns corresponding sub-concepts in their individual subspaces. To model intricate contextuality between sub-concepts and their visual features, compositions are generated from these subspaces in three hierarchical forms, and the composed concepts are learned in a unified composition space. To further refine the captured contextual relationships, adaptively semi-positive concepts are defined and then learned with pseudo supervision exploited from the generated compositions. We validate the proposed approach on two challenging benchmarks, and demonstrate its superiority over state-of-the-art approaches.

1. Introduction

A character of human intelligence is the composing ability towards individual concepts [1]. Imagining *big*, it is common for us to come up with *compositional concepts* such as *big building* and *big cat*. The term *big* here is no more an independent single concept, but is a *sub-concept* that can be combined into new concepts with other sub-concepts, reflecting the ability of *compositional generalization*. However, compositional generalization remains an insurmountable obstacle to machines. In this paper, we explore compositional generalization under the zero-shot learning (ZSL) setting, where a model needs to recognize unseen images composed from seen sub-concepts.

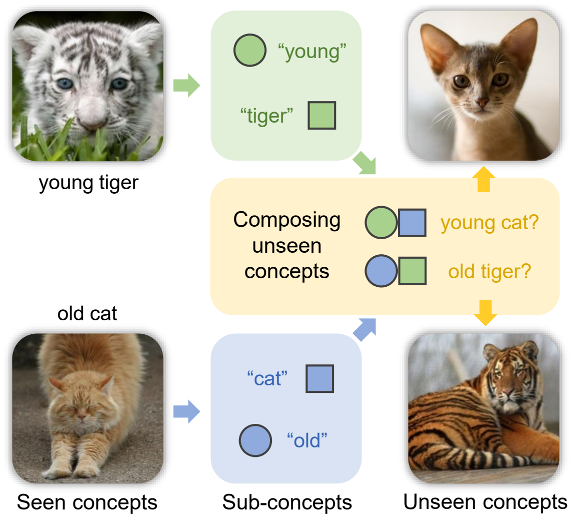

Figure 1. An example of unseen compositional concept recognition. Training with *young tiger* and *old cat*, it is expected to learn the sub-concepts of *young*, *tiger*, *old*, and *cat* that can be used to compose and recognize unseen *young cat* and *old tiger*.

As shown in Figure 1, one main challenge is the *diversity* of sub-concepts, *e.g.*, *young tiger*, where *young* is a semantic description while *tiger* is a physical entity. Sub-concepts can be visually and semantically distinct, resulting in recognition difficulties when facing many possible compositions. Another challenge is the *contextuality* of sub-concepts, *e.g.*, the *old* in *old tiger* should be totally different from the *old* in *old car*. The semantic meanings of sub-concepts tend to highly depend on each other. Contextuality is also associated with specific images [22], *e.g.*, how *old* appears in two different *old tiger* images depends on the images themselves, which requires capturing fine-

*Corresponding author.

grained contextual relationships between sub-concepts and their corresponding visual features. However, most of the recent approaches [17, 18, 27] degrade the contextuality to rigid semantic relationships in a common embedding space, in which images and corresponding word vectors are embedded. They fail to capture richer fine-grained contextual relationships and suffer from intricate contextuality under challenging ZSL setting.

In this paper, we propose to learn unseen compositional concepts in a hierarchical decomposition-and-composition (HiDC) manner. Specifically, to address the challenge of *diversity*, HiDC decomposes each seen image into visual elements according to its compositional labels, and learns the decomposed sub-concepts in their individual subspaces. For *contextuality*, various compositions are generated from these two concept subspaces, and the composed concepts are learned in a unified composition space. To further explore intricate contextuality between sub-concepts and their corresponding visual features, we propose to generate compositions in three hierarchical forms, *i.e.*, *visual*, *word*, and *hybrid* compositions. Visual compositions serve as visual prototypes for concept learning; word compositions emulate concept learning by mapping word concepts to visual features; hybrid compositions bridge visual and word concepts by transmitting visual features to word concepts and vice versa. These three hierarchical compositions are able to model intricate contextuality, and allow capturing fine-grained contextual relationships. Moreover, to refine the contextuality captured by the compositions, we discover and compose *adaptively semi-positive* concepts. To this end, underlying knowledge of the generated compositions is exploited as adaptive pseudo supervision to learn the semi-positive concepts more accurately. Our proposed HiDC is validated on two popular benchmark datasets. Experiments demonstrate that HiDC consistently outperforms the state of the arts. Also, the ablation study verifies the effectiveness of each proposed module.

To sum up, the main contributions of this paper are:

- An end-to-end decomposition-and-composition approach with three hierarchical composition forms to model intricate contextuality between compositional sub-concepts and their corresponding visual features;
- A novel exploration of *adaptively semi-positive* concepts that depict fine-grained contextual relationships; also, an exploitation of adaptive pseudo supervision from the generated compositions to learn such semi-positive concepts accurately;
- Extensive ablation studies and experiments, which validate the effectiveness of our proposed approach and demonstrate its superiority over the state of the arts.

2. Related Work

Zero-Shot Learning (ZSL). The aim of ZSL [20, 14, 15] is transferring knowledge from seen concepts to unseen ones, such that a model is able to recognize new concepts which never appear in training. Basically, mainstream ZSL methods can be divided into two categories: 1) embedding-based methods and 2) generating-based methods. Embedding-based methods [23, 2, 3, 28, 33] aim to find a discriminative common embedding space for both visual features and attribute semantic features. Generating-based methods [13, 26, 10, 19, 31] utilize generative models to synthesize unseen concepts. ZSL can be further extended to a more practical setting, *i.e.*, generalized ZSL (GZSL), where the models are required to identify an unseen concept with a seen/unseen label. By contrast, conventional ZSL only requires to identify an unseen concept with an unseen label. In this paper, we propose an embedding-based GZSL method where hierarchical embedding spaces are constructed to learn compositional concepts.

Unseen Compositional Concept Recognition. This task is a specialized ZSL problem where images are labeled with compositional concepts, *e.g.*, *young tiger*. Early methods [4, 17, 25] often train independent classifiers for each sub-concept, and combine the trained classifiers to recognize unseen concepts. Methods most relevant to ours are *AttrAsOperator* [18] and *AdvFineGrained* [27]. AttrAsOperator regards the compositional concepts in the datasets as attribute-object pairs. By treating attributes (*e.g.*, *young*) as operators, AttrAsOperator composes attribute-conditioned transformations in a common embedding space to learn unseen attribute-object concepts with triplet loss regularization [9, 5, 8, 29, 30]. In contrast, rather than treating compositional concepts as attribute-object pairs, we decompose and compose these concepts in a unified framework without explicitly modeling each sub-concept. This enables generalizability in real-world applications when compositions are not biased to attribute-object pairs. On the other hand, AdvFineGrained proposes to regulate the common embedding space with a quintuplet loss, where *semi-negative* samples are defined. We argue that this definition is too rigid to learn accurate compositional concepts (see *Word Compositions* in Section 3.2 for detailed discussion). On the contrary, we regard them as *adaptively semi-positive* samples to learn more accurate concepts. Additionally, AdvFineGrained employs multi-scale features and adversarial training for better performance; while we do not involve any of the problem-unrelated tricks and still achieve superior performance.

3. Approach

We consider the setting where each image $\mathbf{I}_{a,o}$ is consisted of an attribute y_a and an object y_o, and its label y is

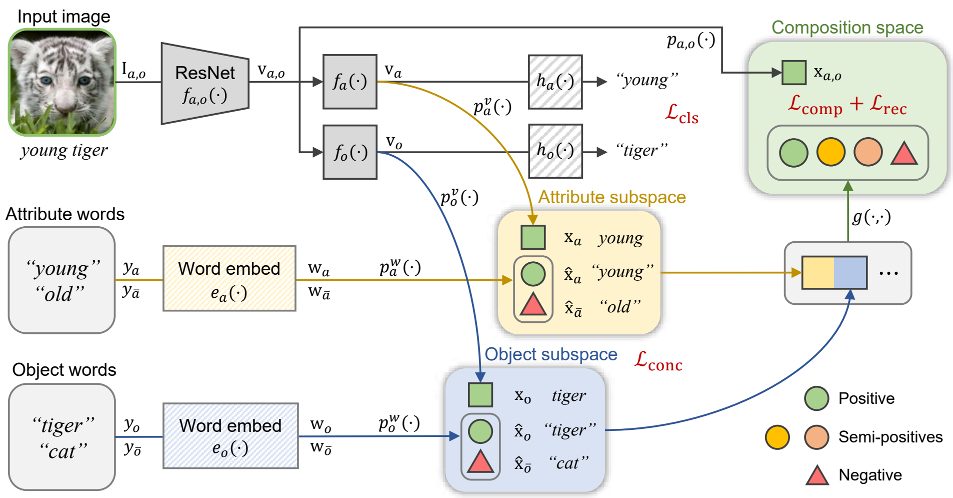

Figure 2. Pipeline of the proposed HiDC, which mainly includes decomposing visual features into attribute and object subspaces, and generating hierarchical compositions from these subspaces. Best viewed in color.

denoted as $y = (y_a, y_o)$. Note that we only formulate the compositional concepts as attribute-object pairs for brevity, but do not explicitly model each sub-concept as in AttrAsOperator [18]. As a typical ZSL task, given an unseen image $\mathbf{I}$, the goal is to predict its corresponding label (y_a, y_o). To this end, the dataset is divided into two parts, *i.e.*, seen part $\mathcal{D}^s = \{(\mathbf{I}_{a,o}, y) \mid \mathbf{I}_{a,o} \in \mathcal{X}^s, y \in \mathcal{Y}^s\}$ for training, and unseen part $\mathcal{D}^u = \{(\mathbf{I}, y) \mid \mathbf{I} \in \mathcal{X}^u, y \in \mathcal{Y}^u\}$ for test, where $\mathcal{Y} = \mathcal{Y}^s \cup \mathcal{Y}^u = \mathcal{Y}_a \times \mathcal{Y}_o = \{(y_a, y_o) \mid y_a \in \mathcal{Y}_a, y_o \in \mathcal{Y}_o\}$. The training and test labels are non-overlapping: $\mathcal{Y}^s \cap \mathcal{Y}^u = \varnothing$. Under this setting, all attributes and objects can be seen during training: $\mathcal{Y}^u_a \subseteq \mathcal{Y}^s_a, \mathcal{Y}^u_o \subseteq \mathcal{Y}^s_o$ where $y_a \in \mathcal{Y}_a$ and $y_o \in \mathcal{Y}_o$. In other words, all the attribute/object sub-concepts are available for training, but the composed attribute-object concepts are non-overlapping for training and test. For GZSL, we expect to learn a prediction $\mathcal{X}^u \mapsto \mathcal{Y}^u \cup \mathcal{Y}^s$ by training on $\{\mathcal{X}^s, \mathcal{Y}^s\}$.

In the following subsections, we will introduce the components of our proposed HiDC, followed by the description of training and test procedures.

3.1. Concept Decomposition

Decomposing Visual Features. Given an image $\mathbf{I}_{a,o}$ labeled with (y_a, y_o), we first input it to a pre-trained ResNet-18 [7] to extract high-level visual features as $\mathbf{v}_{a,o} = f_{a,o}(\mathbf{I}_{a,o})$. As illustrated in Figure 2, the extracted features $\mathbf{v}_{a,o}$ are directly fed into two individual MLPs as $\mathbf{v}_a = f_a(\mathbf{v}_{a,o})$ and $\mathbf{v}_o = f_o(\mathbf{v}_{a,o})$, each followed by a separate classifier. The classifiers are used to predict attribute label y_a and object label y_o of $\mathbf{v}_a$ and $\mathbf{v}_o$ respectively. The classification loss is written as

$$\mathcal{L}_{\text{cls}}(\mathbf{v}_a, \mathbf{v}_o) = h_a(\mathbf{v}_a, y_a) + h_o(\mathbf{v}_o, y_o)\,, \qquad (1)$$

where $h_a(\cdot)$ and $h_o(\cdot)$ denote two fully-connected layers, each containing a cross-entropy loss trained to classify attributes and objects respectively.

Constructing Concept Subspaces. The two output vectors $\mathbf{v}_a$ and $\mathbf{v}_o$ are projected into two embedding subspaces, *i.e.*, attribute and object subspaces, as $\mathbf{x}_a = p^v_a(\mathbf{v}_a)$ and $\mathbf{x}_o = p^v_o(\mathbf{v}_o)$. Meanwhile, the corresponding attribute and object are embedded as word vectors $\mathbf{w}_a = e_a(y_a)$ and $\mathbf{w}_o = e_o(y_o)$. We also randomly embed an attribute word as $\mathbf{w}_{\bar{a}}$ and an object word as $\mathbf{w}_{\bar{o}}$ different from $\mathbf{w}_a$ and $\mathbf{w}_o$, where "$^-$" is a negative index. Then the attribute and object word vectors $\mathbf{w}_a$, $\mathbf{w}_{\bar{a}}$, $\mathbf{w}_o$ and $\mathbf{w}_{\bar{o}}$ are projected into the attribute and object subspaces as positive samples $\hat{\mathbf{x}}_a = p^w_a(\mathbf{w}_a), \hat{\mathbf{x}}_o = p^w_o(\mathbf{w}_o)$ and negative samples $\hat{\mathbf{x}}_{\bar{a}} = p^w_a(\mathbf{w}_{\bar{a}}), \hat{\mathbf{x}}_{\bar{o}} = p^w_o(\mathbf{w}_{\bar{o}})$. Together with the anchors $\mathbf{x}_a$ and $\mathbf{x}_o$, we can construct two triplets in these two concept subspaces, which are regularized by two triplet losses. The triplet loss pulls positive samples close to the anchor, and pushes the negative ones away from the anchor. The triplet losses on the two concept subspaces are defined as

$$\mathcal{L}^a_{\text{triplet}}(\mathbf{x}_a, \hat{\mathbf{x}}_a, \hat{\mathbf{x}}_{\bar{a}}) = \max\left(0, d(\mathbf{x}_a, \hat{\mathbf{x}}_a) - d(\mathbf{x}_a, \hat{\mathbf{x}}_{\bar{a}}) + m\right),$$
$$\mathcal{L}^o_{\text{triplet}}(\mathbf{x}_o, \hat{\mathbf{x}}_o, \hat{\mathbf{x}}_{\bar{o}}) = \max\left(0, d(\mathbf{x}_o, \hat{\mathbf{x}}_o) - d(\mathbf{x}_o, \hat{\mathbf{x}}_{\bar{o}}) + m\right),$$

where $d(\cdot, \cdot)$ denotes Euclidean distance, and m is the triplet margin value, the same as below. The overall loss on the two concept subspaces are added by the two triplet losses:

$$\mathcal{L}_{\text{conc}} = \mathcal{L}^a_{\text{triplet}} + \mathcal{L}^o_{\text{triplet}}\,. \qquad (2)$$

3.2. Hierarchical Concept Composition

With the two constructed concept subspaces, we can flexibly create and adjust attribute-object compositions. We implement the composition by concatenating attribute and

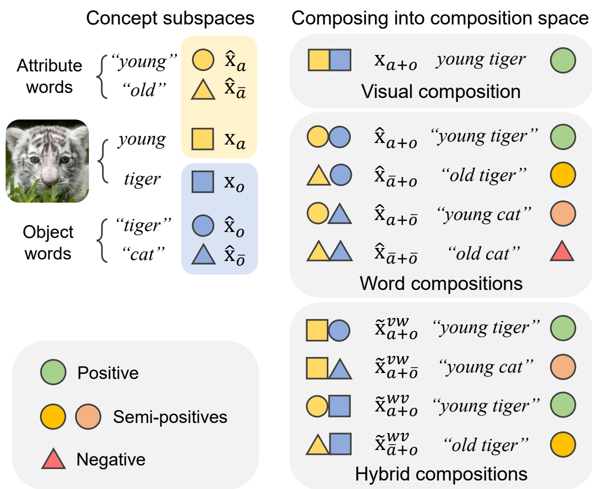

Figure 3. Toy illustration of the three hierarchical forms of compositions, where an attribute element and an object one are taken from each of the concept subspaces for composing.

object elements from the two concept subspaces, and sending them to a fully-connected layer. As shown in Figure 3, the compositions are divided into three hierarchical forms, *i.e.*, *Visual*, *Word*, and *Hybrid* Compositions.

Visual Composition. Visual embeddings $\mathbf{x}_a$ and $\mathbf{x}_o$ in the two concept subspaces are directly combined to obtain a visual composition as $\mathbf{x}_{a+o} = g(\mathbf{x}_a, \mathbf{x}_o)$. Naturally, the visual composition should resemble the original image. As shown in Figure 2, the output vector $\mathbf{v}_{a,o}$ of the pre-trained feature extractor is embedded into the composition space as a universal anchor $\mathbf{x}_{a,o} = p_{a,o}(\mathbf{v}_{a,o})$ for all generated compositions. The visual composition $\mathbf{x}_{a+o}$ is regularized to reconstruct the image embedding $\mathbf{x}_{a,o}$:

$$\mathcal{L}_{\text{rec}}(\mathbf{x}_{a+o}, \mathbf{x}_{a,o}) = d(\mathbf{x}_{a+o}, \mathbf{x}_{a,o})\,. \tag{3}$$

Word Compositions. There are four word compositions, *i.e.*, $\hat{\mathbf{x}}_{a+o} = g(\hat{\mathbf{x}}_a, \hat{\mathbf{x}}_o)$, $\hat{\mathbf{x}}_{\bar{a}+o} = g(\hat{\mathbf{x}}_{\bar{a}}, \hat{\mathbf{x}}_o)$, $\hat{\mathbf{x}}_{a+\bar{o}} = g(\hat{\mathbf{x}}_a, \hat{\mathbf{x}}_{\bar{o}})$, and $\hat{\mathbf{x}}_{\bar{a}+\bar{o}} = g(\hat{\mathbf{x}}_{\bar{a}}, \hat{\mathbf{x}}_{\bar{o}})$. Covering all the unseen compositional concepts, word compositions are the main concerns to unseen concept learning.

A recent study [27] proposes to treat the compositions like $\hat{\mathbf{x}}_{\bar{a}+o}$ and $\hat{\mathbf{x}}_{a+\bar{o}}$ as *semi-negative* samples, supplementary to the conventional positive and negative compositions $\hat{\mathbf{x}}_{a+o}, \hat{\mathbf{x}}_{\bar{a}+\bar{o}}$. The problem is "semi-negative" itself—rigidly treating $\hat{\mathbf{x}}_{\bar{a}+o}$ and $\hat{\mathbf{x}}_{a+\bar{o}}$ as somewhat negative is disadvantageous to compositional concepts learning. In contrast, we treat them as *adaptively semi-positive* samples with variable triplet margins, and formulate the composition loss as

$$\begin{aligned}
&\mathcal{L}_{\text{comp}}(\mathbf{x}_{a,o}, \hat{\mathbf{x}}_{a+o}, \hat{\mathbf{x}}_{\bar{a}+o}, \hat{\mathbf{x}}_{a+\bar{o}}, \hat{\mathbf{x}}_{\bar{a}+\bar{o}}) = \\
&\quad \max\big(0, d(\mathbf{x}_{a,o}, \hat{\mathbf{x}}_{a+o}) - d(\mathbf{x}_{a,o}, \hat{\mathbf{x}}_{\bar{a}+\bar{o}}) + m\big) \\
&+ \max\big(0, d(\mathbf{x}_{a,o}, \hat{\mathbf{x}}_{\bar{a}+o}) - d(\mathbf{x}_{a,o}, \hat{\mathbf{x}}_{\bar{a}+\bar{o}}) + \alpha m\big) \\
&+ \max\big(0, d(\mathbf{x}_{a,o}, \hat{\mathbf{x}}_{a+\bar{o}}) - d(\mathbf{x}_{a,o}, \hat{\mathbf{x}}_{\bar{a}+\bar{o}}) + \beta m\big)\,,
\end{aligned} \tag{4}$$

where we introduce $\alpha, \beta \in (0, 1)$ as margin adjusting parameters. The motivation is from a simple observation: if given *young tiger* as an anchor, *young cat* should be closer to the anchor than *young horse* does, where both *young cat* and *young horse* are adaptively regarded as semi-positive samples. In Eq. (4), a smaller triplet margin pulls semi-positive samples $\hat{\mathbf{x}}_{\bar{a}+o}, \hat{\mathbf{x}}_{a+\bar{o}}$ close to the negative one $\hat{\mathbf{x}}_{\bar{a}+\bar{o}}$, and meanwhile pushes them away from the anchor $\mathbf{x}_{a,o}$.

Thus we fix the triplet margin of the positive sample $\hat{\mathbf{x}}_{a+o}$ and assign each semi-positive sample $\hat{\mathbf{x}}_{\bar{a}+o}, \hat{\mathbf{x}}_{a+\bar{o}}$ an adaptive margin that can be controlled by α and β respectively. The margin adjusting parameters are determined by the underlying knowledge of hybrid compositions introduced below.

Hybrid Compositions. Each composition in this form is generated from a visual embedding and a word one. Hence hybrid compositions are restricted to either positive samples $\tilde{\mathbf{x}}^{vw}_{a+o} = g(\mathbf{x}_a, \hat{\mathbf{x}}_o), \tilde{\mathbf{x}}^{wv}_{a+o} = g(\hat{\mathbf{x}}_a, \mathbf{x}_o)$ or semi-positive ones $\tilde{\mathbf{x}}^{vw}_{a+\bar{o}} = g(\mathbf{x}_a, \hat{\mathbf{x}}_{\bar{o}}), \tilde{\mathbf{x}}^{wv}_{\bar{a}+o} = g(\hat{\mathbf{x}}_{\bar{a}}, \mathbf{x}_o)$, where "$vw$" denotes *visual-word* compositions while "wv" *word-visual* ones.

Let us consider a practical example with an anchor concept as *young tiger* and semi-positive concepts as *young cat* and *young horse*. As shown in Figure 4, the corresponding positive word/hybrid compositions are denoted as $\hat{\mathbf{x}}_{\text{young+tiger}}, \tilde{\mathbf{x}}^{vw}_{\text{young+tiger}}$, and the semi-positive word/hybrid compositions as $\hat{\mathbf{x}}_{\text{young+cat}}, \tilde{\mathbf{x}}^{vw}_{\text{young+cat}}$ and $\hat{\mathbf{x}}_{\text{young+horse}}, \tilde{\mathbf{x}}^{vw}_{\text{young+horse}}$. Since the visual element *young* is decomposed from the image $\mathbf{I}_{\text{young,tiger}}$, hybrid compositions $\tilde{\mathbf{x}}^{vw}_{\text{young+tiger}}$, $\tilde{\mathbf{x}}^{vw}_{\text{young+cat}}$, and $\tilde{\mathbf{x}}^{vw}_{\text{young+horse}}$ should all contain a *tiger*-style *young* as shown in Figure 4. The inconsistency between the visual *tiger*-style *young* and the negative word *cat/horse* actually reflects how much the negative word sub-concept *cat/horse* violates the positive visual sub-concept *young* which is biased towards *tiger*. We propose to measure the inconsistency by calculating the distance between the hybrid composition and its corresponding word composition, and further employ it as adaptive pseudo supervision for margin adjusting, *i.e.*, as illustrated in Figure 4, $d_1 = d(\tilde{\mathbf{x}}^{vw}_{\text{young+cat}}, \hat{\mathbf{x}}_{\text{young+cat}})$, and $d_2 = d(\tilde{\mathbf{x}}^{vw}_{\text{young+horse}}, \hat{\mathbf{x}}_{\text{young+horse}})$. We can expect a larger d_2 than d_1 since a *horse*-style *young* lies farther from a *tiger*-style *young* than a *cat*-style *young* does.

Now we consider the margin adjusting parameter problem. Following the above example, we calculate $d_0 = d(\tilde{\mathbf{x}}^{vw}_{\text{young+tiger}}, \hat{\mathbf{x}}_{\text{young+tiger}})$ as a benchmark distance to d_1 and d_2. When choosing $\hat{\mathbf{x}}_{\text{young+cat}}$ as the semi-positive sample, the margin adjusting parameter is given by $\beta_1 = \sigma(d_0 - d_1)$, where $\sigma(\cdot)$ denotes a sigmoid function. As to $\hat{\mathbf{x}}_{\text{young+horse}}$, we have $\beta_2 = \sigma(d_0 - d_2)$. When $d_1 < d_2$, we can derive $\beta_1 > \beta_2$. As illustrated in Figure 4, the larger β_1 pulls $\hat{\mathbf{x}}_{\text{young+cat}}$ closer to the anchor $\mathbf{x}_{\text{young,tiger}}$ and pushes it away from the negative composition ($\hat{\mathbf{x}}_{\text{old+cat}}$, in this example). In general, the margin adjusting parameters

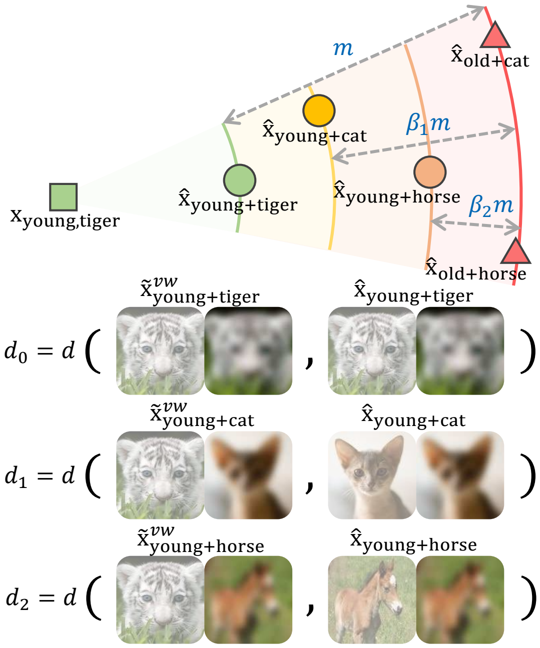

Figure 4. Upper: illustration of margin adjusting strategy, where triplets corresponding to different negative samples are put in the same coordinate for brevity. Lower: conceptual illustration of pseudo supervision exploitation from hybrid compositions, where *young* is artificially separated from *tiger/cat/horse* for clearer illustration.

α and β in Eq. (4) are calculated as

$$\alpha = \sigma\Big(\lambda\big(d(\tilde{\mathbf{x}}^{wv}_{a+o}, \hat{\mathbf{x}}_{a+o}) - d(\tilde{\mathbf{x}}^{wv}_{\bar{a}+o}, \hat{\mathbf{x}}_{\bar{a}+o})\big)\Big)\,, \tag{5}$$

$$\beta = \sigma\Big(\lambda\big(d(\tilde{\mathbf{x}}^{vw}_{a+o}, \hat{\mathbf{x}}_{a+o}) - d(\tilde{\mathbf{x}}^{vw}_{a+\bar{o}}, \hat{\mathbf{x}}_{a+\bar{o}})\big)\Big)\,, \tag{6}$$

where $\sigma(\cdot)$ denotes a sigmoid function and λ scales its input to adjust the sensitive region of the sigmoid function.

3.3. Training and Test

Training. The training procedure is summarized in Algorithm 1, in which image feature decomposition is guided by $\mathcal{L}_{\mathrm{cls}}$ in Eq. (1), the two concept subspaces are constrained by $\mathcal{L}_{\mathrm{conc}}$ in Eq. (2), and the compositions are regularized by $\mathcal{L}_{\mathrm{rec}}$ and $\mathcal{L}_{\mathrm{comp}}$ in Eqs. (3) (4), where α and β in $\mathcal{L}_{\mathrm{comp}}$ are given by Eqs. (5) (6).

Test. The trained model is tested on the unseen set $\mathcal{D}^u$. Given an unseen image $\mathbf{I}$ for test, we first extract and map its visual features to the composition space as $\mathbf{x} = p_{a,o}\big(f_{a,o}(\mathbf{I})\big)$, and then generate word compositions $\hat{\mathbf{x}}_{a+o} = g\Big(p^w_a\big(e_a(y_a)\big), p^w_o\big(e_o(y_o)\big)\Big)$ from all n candidate attribute-object pairs $\{(y_a, y_o)\}^n_{i=1}$. We compute and store the distances between $\mathbf{x}$ and each $\hat{\mathbf{x}}_{a+o}$, and select the label $y = (y_a, y_o)$ of $\hat{\mathbf{x}}_{a+o}$ corresponding to the shortest distance as the prediction of $\mathbf{I}$:

$$P(\mathbf{I}) = \arg\min_{y \in \widetilde{\mathcal{Y}}} d(\mathbf{x}, \hat{\mathbf{x}}_{a+o})\,, \tag{7}$$

where $\widetilde{\mathcal{Y}} = \mathcal{Y}^u$ for conventional ZSL, and $\widetilde{\mathcal{Y}} = \mathcal{Y}^u \cup \mathcal{Y}^s$ for GZSL.

Algorithm 1: Training procedure of hierarchical decomposition and composition (HiDC) model for unseen concept recognition.

```
Data: Training data D^s, scale parameter λ
Result: Optimal e_a, e_o, p^w_a, p^w_o, p_{a,o}, g
Initialize: f_a, f_o, h_a, h_o, e_a, e_o, p^v_a, p^v_o, p^w_a, p^w_o, p_{a,o}, g;
while not converged do
    Sample a batch from D^s as {I_{a,o}}^n_{i=1} with labels
      {(y_a, y_o)}^n_{i=1}; sample corresponding negative
      labels {(y_ā, y_ō)}^n_{i=1} randomly;
    for samples in the batch do
        Decompose image features:
          v_a = f_a(f_{a,o}(I_{a,o})), v_o = f_o(f_{a,o}(I_{a,o}));
        Construct concept subspaces:
          x_a = p^v_a(v_a), x_o = p^v_o(v_o),
          x̂_a = p^w_a(e_a(y_a)), x̂_o = p^w_o(e_o(y_o)),
          x̂_ā = p^w_a(e_a(y_ā)), x̂_ō = p^w_o(e_o(y_ō));
        Generate compositions from concept subspaces:
          x_{a,o} = p_{a,o}(f_{a,o}(I_{a,o})), x_{a+o} = g(x_a, x_o),
          x̂_{a+o} = g(x̂_a, x̂_o), x̂_{ā+o} = g(x̂_ā, x̂_o),
          x̂_{a+ō} = g(x̂_a, x̂_ō), x̂_{ā+ō} = g(x̂_ā, x̂_ō),
          x̃^{vw}_{a+o} = g(x_a, x̂_o), x̃^{wv}_{a+o} = g(x̂_a, x_o),
          x̃^{vw}_{a+ō} = g(x_a, x̂_ō), x̃^{wv}_{ā+o} = g(x̂_ā, x_o);
        Calculate α, β, L_cls, L_conc, L_rec, L_comp by
          Eqs. (5) (6) (1) (2) (3) (4), respectively;
    end
    L_train = L_cls + L_conc + L_rec + L_comp;
    Update network parameters using ∇L_train;
end
```

4. Experiments

In this section, we present ablation studies and parameter analysis to validate the effectiveness of the proposed HiDC, and compare HiDC with several state of the arts to verify its superiority over current methods.

4.1. Experimental Setup

Datasets. Our proposed HiDC and baselines are evaluated on two popular benchmark datasets, *i.e.*, *MIT-States* [11] and *UT-Zappos* [32].

MIT-States contains 53,753 everyday images with a wide range of attributes (115 classes) and objects (245 classes). Each image is annotated with an attribute-object concept such as "young tiger", and there are 1962 pairs in total. We use the compositional split [17], *i.e.*, 1262 pairs in $\mathcal{Y}^s$ for training and 700 pairs in $\mathcal{Y}^u$ for test.

Dataset	Attribute Classes	Object Classes	Seen Pairs (Images)	Unseen Pairs (Images)
MIT-States [11]	115	245	1262 (34,562)	700 (19,191)
UT-Zappos [32]	16	12	83 (24,898)	33 (4,228)

Table 1. Dataset descriptions with the numbers of attribute/object classes and seen/unseen pairs (images).

#	Modules					MIT-States			UT-Zappos		
	$\mathcal{L}_{\text{cls}}$	$\mathcal{L}_{\text{conc}}$	$\mathcal{L}_{\text{comp}}$	$\mathcal{L}_{\text{rec}}$	α, β	Closed	Open	H-Mean	Closed	Open	H-Mean
1	✓					12.6	2.3	3.9	39.0	4.8	8.6
2	✓	✓				13.3	2.5	4.2	42.1	5.3	9.5
3	✓	✓	✓			15.0	12.3	13.5	50.6	47.2	48.8
4	✓	✓	✓	✓		14.7	13.0	13.8	52.3	47.3	49.7
5	✓	✓	✓		✓	15.2	14.3	14.7	52.4	51.5	52.0
6	✓	✓	✓	✓	✓	**15.4**	**14.6**	**15.0**	**53.4**	**51.5**	**52.4**

Table 2. Ablation study on the five proposed modules. Results are reported in unseen pair recognition accuracy (%) under three evaluation metrics on the two datasets. Note that ✓under α, β denotes that the margin adjusting strategy is enabled.

UT-Zappos contains 50,025 images of shoes, where each image is annotated with an attribute-object concept such as "canvas slippers". There are 16 attribute classes and 12 object classes. Following the same setting in [18, 27], we use the subset of 29,126 images in the experiments, *i.e.*, 83 attribute-object pairs in $\mathcal{Y}^s$ for training and 33 pairs in $\mathcal{Y}^u$ for test. Table 1 summaries the details of the two datasets.

Evaluation Metrics. We follow the same evaluation standard in [18, 27]. The top-1 accuracy of unseen attribute-object concept recognition is reported under three metrics:

1) *Closed*, where the test candidate attribute-object pairs are from $\mathcal{Y}^u$. Closed metric evaluates the recognition ability on unseen concepts. This metric restricts the test candidates into unseen pairs, in most cases yields higher accuracy due to reduced number of test candidates.

2) *Open*, where the test candidate attribute-object pairs are from $\mathcal{Y}^u \cup \mathcal{Y}^s$, corresponding to GZSL setting. Open metric evaluates the general recognition ability on both seen and unseen concepts. This metric often yields a relatively lower accuracy since all seen and unseen attribute-object pairs are included into the test candidates, and thus is more practical for real-world applications.

3) *H-Mean*, namely *harmonic mean*, which consolidates both Closed and Open metrics. H-Mean is defined as

$$A_{\text{H}} = 2 \times \frac{A_{\text{Closed}} \times A_{\text{Open}}}{A_{\text{Closed}} + A_{\text{Open}}}, \tag{8}$$

where A_{H}, A_{Closed}, and A_{Open} are the accuracy measured under H-Mean, Closed, and Open metrics respectively. H-Mean penalizes large performance discrepancy between Closed and Open metrics, which has been widely adopted in GZSL for evaluating the overall generalizability.

Implementation Details. Visual feature extractor $f_{a,o}(\cdot)$ is implemented as ResNet-18 [7] pre-trained on ImageNet [24], without fine-tuning for fair comparison with all the baselines. Projections $p(\cdot)$ are implemented as fully-connected layers, and the same with $g(\cdot)$. Also, word embeddings $e(\cdot)$ are fully-connected layers trained from scratch without extra knowledge. The triplet margin m is set to 2. Our model is implemented in PyTorch (version 1.1.0) with ADAM [12] optimizer. The code will be made public available.

4.2. Ablation Study and Analysis

Loss Functions. We ablate our model to evaluate the effectiveness of the proposed modules:

1) *Base model*, which is only composed of the MLPs and the followed classifiers. The training is guided by $\mathcal{L}_{\text{cls}}$, and we test it by directly predicting attribute/object labels of test set images with the classifiers.

2) *Adding* $\mathcal{L}_{\text{conc}}$, which further incorporates the attribute/object subspace constraints. The training is guided by $\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{conc}}$, and we test it the same way as above.

3) *Adding* $\mathcal{L}_{\text{comp}}$ *(without margin adjusting)*. The training is guided by $\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{conc}} + \mathcal{L}_{\text{comp}}$. Margin adjusting parameters α and β in $\mathcal{L}_{\text{comp}}$ are both fixed to 0.5. We test it as described in Section 3.3, the same as below.

4) *Adding* $\mathcal{L}_{\text{rec}}$. The training is guided by $\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{conc}} + \mathcal{L}_{\text{comp}} + \mathcal{L}_{\text{rec}}$, where α and β still remain 0.5.

5) *Adding α and β (without $\mathcal{L}_{\text{rec}}$)*. The training is guided by $\mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{conc}} + \mathcal{L}_{\text{comp}}$, where α and β are now variables given by Eqs. (5) (6).

6) *Full model.*

As shown in Table 2, each of the proposed modules contributes to the overall performance. Compared with Closed metric, our base model performs considerably worse under Open metric, for its lacking of the capability to learn contextual relationships between compositional sub-concepts, and thus fails to transfer knowledge from seen concepts to unseen ones. In contrast, our proposed $\mathcal{L}_{\text{comp}}$ models contextual relationships by regularizing compositions generated from sub-concepts, where $\mathcal{L}_{\text{conc}}$ ensures robust concept subspaces and $\mathcal{L}_{\text{rec}}$ benefits the composing ability. Together with the adaptive learning ability offered by α and β,

Method	MIT-States					UT-Zappos				
	Closed	Open	H-Mean	Attribute	Object	Closed	Open	H-Mean	Attribute	Object
HiDC (with $\mathcal{L}_{\text{quin}}$)	14.6	12.2	13.3	20.4	25.8	52.7	47.1	49.7	51.1	77.4
HiDC (with $\mathcal{L}_{\text{comp}}$)	**15.4**	**14.6**	**15.0**	**22.6**	**26.9**	**53.4**	**51.5**	**52.4**	**55.8**	**77.6**

Table 3. Comparison between the proposed $\mathcal{L}_{\text{comp}}$ in Eq. (4) and $\mathcal{L}_{\text{quin}}$ in [27]. Results are reported in unseen pair recognition accuracy (%) under three evaluation metrics on the two datasets. Also the attribute/object recognition accuracy (%) is reported as complementary.

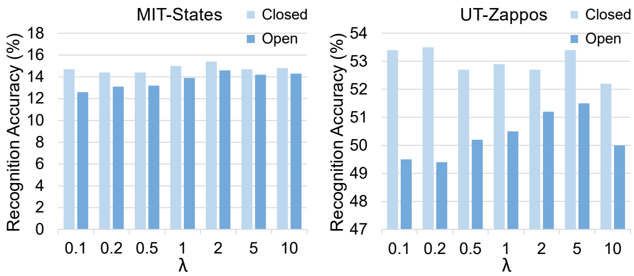

Figure 5. Analysis of scale parameter λ. Results are reported in unseen pair recognition accuracy (%) with changing λ.

our full model achieves favorable performance under both Closed and Open metrics. Arguably, adding trade-off parameters to each loss may benefit the accuracy, but we do not involve any trade-offs for generality and practicality of the proposed HiDC.

Effect of Margin Adjusting. To further verify our adaptive learning ability, we compare the proposed $\mathcal{L}_{\text{comp}}$ with the quintuplet loss $\mathcal{L}_{\text{quin}}$ in current state of the art [27], where we implement our proposed HiDC with $\mathcal{L}_{\text{quin}}$ in replace of $\mathcal{L}_{\text{comp}}$. As shown in Table 3, our proposed $\mathcal{L}_{\text{comp}}$ is comprehensively superior to $\mathcal{L}_{\text{quin}}$ as $\mathcal{L}_{\text{comp}}$ improves the overall recognition performance, especially under the challenging Open metric. Moreover, benefiting from the adaptive learning ability, $\mathcal{L}_{\text{comp}}$ is able to capture fine-grained attribute concepts more accurately than $\mathcal{L}_{\text{quin}}$.

Effect of Scale Parameter λ. The scale parameter λ is employed to control the input scale of the sigmoid function for calculating α and β in Eqs. (5) (6), which can be of help when using different datasets. Roughly, the sensitive region of a sigmoid function is $[-5, 5]$, and the order of magnitude of its input in our experiments is around -1. We select λ in $\{0.1, 0.2, 0.5, 1, 2, 5, 10\}$ and report the recognition accuracy in Figure 5. Optimal performances can be observed when λ is set to around 2. A suitable λ is able to decrease the performance discrepancy between Closed and Open metrics, contributing to the generalizability from seen concepts to unseen ones.

4.3. Comparisons with State of the Arts

Baseline Methods. We compare our proposed HiDC against seven baselines:

1) *VisProd* [16] trains two independent linear SVMs to predict attributes and objects.

2) *AnalogousAttr* [4] trains a linear SVM for each seen pair and generates classifier weights for unseen pairs with Bayesian Probabilistic Tensor Factorization (BPTF).

3) *RedWine* [17] trains linear SVMs for attribute/object sub-concepts and transforms the SVM weights with a neural network for unseen pairs.

4) *LabelEmbed* [6] uses pre-trained GloVe [21] word embeddings rather than classifier weights to compose word vector representations, compared with *RedWine*.

5) *LabelEmbed+* [18] improves *LabelEmbed* by incorporating image features and training input representations.

6) *AttrAsOperator* [18] treats attributes as operators and conducts attribute-conditioned transformations to learn unseen attribute-object pairs.

7) *AdvFineGrained* [27] defines semi-negative samples and regulates them with a quintuplet loss. For fair comparison, we report its results without using multi-scale features to keep consistency with the other methods.

Quantitative Results. As demonstrated in Table 4, our proposed HiDC consistently outperforms all seven baselines under all evaluation metrics. Except AttrAsOperator and AdvFineGrained, all other baselines perform exceedingly worse under Open metric than Closed, which actually suggests over-fitting to a subset of concepts. In contrast, HiDC exhibits the least performance discrepancy between Closed and Open metrics, verifying its superior generalizability from seen concepts to unseen ones. As we discussed in Section 2, AttrAsOperator unequally models each compositional sub-concept, and thus cannot generalize well on UT-Zappos and performs considerably worse than AdvFineGrained and HiDC that equally treat sub-concepts. Benefiting from our adaptive learning strategy, HiDC is able to capture more accurate fine-grained compositional relationships and outperforms state-of-the-art AdvFineGrained under all metrics. Compared to UT-Zappos, the overall worse performance on MIT-States is due to a larger number of unseen pairs with fewer number of training images for each pair, and also the images are more complicated.

Qualitative Results. Our trained model can be directly employed to retrieve relevant images with text queries given as unseen attribute-object pairs (y_a, y_o). We embed a query (y_a, y_o) and all image candidates $\mathbf{I}$ into the composition space as $\hat{\mathbf{x}}_{a+o}$ and $\mathbf{x}$, and store the distances between $\hat{\mathbf{x}}_{a+o}$ and each $\mathbf{x}$. Corresponding nearest images are selected as results. Figure 6 gives retrieval results of similar concepts that only differ in attributes. Our method outperforms peer

Method	MIT-States			UT-Zappos		
	Closed	Open	H-Mean	Closed	Open	H-Mean
VisProd [16]	11.1	2.4	3.9	46.8	4.1	7.5
AnalogousAttr [4]	1.4	0.2	0.4	18.3	3.5	5.9
RedWine [17]	12.5	3.1	5.0	40.3	2.1	4.0
LabelEmbed [6]	13.4	3.3	5.3	25.8	5.2	8.7
LabelEmbed+ [18]	14.8	5.7	8.2	37.4	9.4	15.0
AttrAsOperator [18]	12.0	11.4	11.7	33.2	23.4	27.5
AdvFineGrained [27]	13.9	12.3	13.1	52.1	48.4	50.2
HiDC (Ours)	**15.4**	**14.6**	**15.0**	**53.4**	**51.5**	**52.4**

Table 4. Comparison between our proposed HiDC and seven baselines. Results are reported in unseen pair recognition accuracy (%) under three evaluation metrics on the two datasets.

Figure 6. Qualitative results of retrieving *young tiger*, *old tiger* in MIT-States, and *sheepskin ankle-boots*, *leather ankle-boots* in UT-Zappos. The top-6 results of AttrAsOperator [18], AdvFineGrained [27], and our proposed HiDC are reported. Correct and incorrect results are respectively marked in green and red borders.

methods [18, 27] and is better at distinguishing similar concepts. Still, attributes such as *old* are poorly learned due to ambiguous visual features and few training images.

5. Conclusions

In this paper, we present a hierarchical decomposition-and-composition (HiDC) model for unseen compositional concept recognition. We propose to decompose each seen image as visual elements and learn the corresponding sub-concepts in independent subspaces. We generate compositions from these subspaces in three hierarchical forms, and learn the composed concepts in a unified composition space. We define semi-positive concepts to depict fine-grained contextual relationships between sub-concepts, and learn accurate compositional concepts with adaptive pseudo supervision exploited from the generated compositions. Extensive ablation studies and experiments validate the effectiveness of our proposed HiDC, and demonstrate its superiority over state-of-the-art approaches. Still, HiDC is limited to compositions with two seen sub-concepts. Extensions to compositions with more sub-concepts (or even unseen ones during training) will be our future work.

Acknowledgment

This research was partially supported by the Key R&D Program-The Key Industry Innovation Chain of Shaanxi under Grant (2018ZDXM-GY-176 and 2019ZDLGY03-02-01), the National Key R&D Program of China under Grant (2017YFE0104100, 2016YFE0200400, 2018AAA0100704, and 2016YFB1001003), NSFC (61972250, U19B2035, U1609220, and 61672231), STCSM (18DZ1112300), and (ARC FL-170100117, DP-180103424) of Australia.

References

[1] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

[2] Soravit Changpinyo, Wei-Lun Chao, and Fei Sha. Predicting visual exemplars of unseen classes for zero-shot learning. In *ICCV*, pages 3476–3485, 2017.

[3] Binghui Chen and Weihong Deng. Hybrid-attention based decoupled metric learning for zero-shot image retrieval. In *CVPR*, pages 2750–2759, 2019.

[4] Chao-Yeh Chen and Kristen Grauman. Inferring analogous attributes. In *CVPR*, pages 200–207, 2014.

[5] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *CVPR*, pages 403–412, 2017.

[6] Mohamed Elhoseiny, Babak Saleh, and Ahmed Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *ICCV*, pages 2584–2591, 2013.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[8] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[9] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *SIMBAD*, pages 84–92, 2015.

[10] He Huang, Changhu Wang, Philip S Yu, and Chang-Dong Wang. Generative dual adversarial network for generalized zero-shot learning. In *CVPR*, pages 801–810, 2019.

[11] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *CVPR*, pages 1383–1391, 2015.

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

[13] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *CVPR*, pages 3174–3183, 2017.

[14] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, pages 951–958, 2009.

[15] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(3):453–465, 2013.

[16] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *ECCV*, pages 852–869, 2016.

[17] Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. In *CVPR*, pages 1792–1801, 2017.

[18] Tushar Nagarajan and Kristen Grauman. Attributes as operators: factorizing unseen attribute-object compositions. In *ECCV*, pages 169–185, 2018.

[19] Jian Ni, Shanghang Zhang, and Haiyong Xie. Dual adversarial semantics-consistent network for generalized zero-shot learning. In *NeurIPS*, 2019.

[20] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. Zero-shot learning with semantic output codes. In *NeurIPS*, pages 1410–1418, 2009.

[21] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

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

[23] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *ICML*, pages 2152–2161, 2015.

[24] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015.

[25] Rodrigo Santa Cruz, Basura Fernando, Anoop Cherian, and Stephen Gould. Neural algebra of classifiers. In *WACV*, pages 729–737, 2018.

[26] Wenlin Wang, Yunchen Pu, Vinay Kumar Verma, Kai Fan, Yizhe Zhang, Changyou Chen, Piyush Rai, and Lawrence Carin. Zero-shot learning via class-conditioned deep generative models. In *AAAI*, 2018.

[27] Kun Wei, Muli Yang, Hao Wang, Cheng Deng, and Xianglong Liu. Adversarial fine-grained composition learning for unseen attribute-object recognition. In *ICCV*, pages 3741–3749, 2019.

[28] Xinyi Xu, Huanhuan Cao, Yanhua Yang, Erkun Yang, and Cheng Deng. Zero-shot metric learning. In *IJCAI*, pages 3996–4002, 2019.

[29] Xinyi Xu, Yanhua Yang, Cheng Deng, and Feng Zheng. Deep asymmetric metric learning via rich relationship mining. In *CVPR*, pages 4076–4085, 2019.

[30] Xu Yang, Cheng Deng, Xianglong Liu, and Feiping Nie. New $\ell_{2,1}$-norm relaxation of multi-way graph cut for clustering. In *AAAI*, 2018.

[31] Xu Yang, Cheng Deng, Feng Zheng, Junchi Yan, and Wei Liu. Deep spectral clustering using dual autoencoder network. In *CVPR*, pages 4066–4075, 2019.

[32] Aron Yu and Kristen Grauman. Fine-grained visual comparisons with local learning. In *CVPR*, pages 192–199, 2014.

[33] Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. Generalized zero-shot recognition based on visually semantic embedding. In *CVPR*, pages 2995–3003, 2019.