ProCC: Progressive Cross-Primitive Compatibility for Open-World Compositional Zero-Shot Learning

Fushuo Huo[1], Wenchao Xu[1*], Song Guo[3], Jingcai Guo[1, 2*], Haozhao Wang[4], Ziming Liu[1], Xiaocheng Lu[3]

[1]Department of Computing, The Hong Kong Polytechnic University, Hong Kong SAR
[2]The Hong Kong Polytechnic University Shenzhen Research Institute, Shenzhen, China
[3]Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR
[4]School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China
{fushuo.huo, ziming.liu}@connect.polyu.hk, songguo@cse.ust.hk, {jc-jingcai.guo, wenchao.xu}@polyu.edu.hk, hz_wang@hust.edu.cn, xiaochenglu1997@gmail.com

Abstract

Open-World Compositional Zero-shot Learning (OW-CZSL) aims to recognize novel compositions of state and object primitives in images with no priors on the compositional space, which induces a tremendously large output space containing all possible state-object compositions. Existing works either learn the joint compositional state-object embedding or predict simple primitives with separate classifiers. However, the former method heavily relies on external word embedding methods, and the latter ignores the interactions of interdependent primitives, respectively. In this paper, we revisit the primitive prediction approach and propose a novel method, termed Progressive Cross-primitive Compatibility (ProCC), to mimic the human learning process for OW-CZSL tasks. Specifically, the cross-primitive compatibility module *explicitly* learns to model the interactions of state and object features with the trainable memory units, which efficiently acquires cross-primitive visual attention to reason high-feasibility compositions, *without* the aid of external knowledge. Moreover, to alleviate the invalid cross-primitive interactions, especially for partial-supervision conditions (pCZSL), we design a progressive training paradigm to optimize the primitive classifiers conditioned on pre-trained features in an easy-to-hard manner. Extensive experiments on three widely used benchmark datasets demonstrate that our method outperforms other representative methods on both OW-CZSL and pCZSL settings by large margins.

Introduction

Humans can extrapolate new concepts from previously learned knowledge. For instance, if the people are taught what the *fried chip* and *toasted bread* are, most of them can recognize the *fried bread* immediately. This ability is known as *compositional generalization* (Atzmon et al. 2016), which is one of the ultimate targets for artificial intelligence. In the literature, such a task is formulated as *Compositional Zero-Shot Learning* (CZSL). Concretely, the training set contains images with corresponding descriptions (primitives), i.e., state and object. The model is expected to recognize unseen compositions based on known primitives, which is non-trivial because object and state are semantically tangled, i.e., objects in different states often have different appearances, and states can vary greatly conditioned on different objects. The major challenge behind the CZSL lies in how to model the interactions between state and object primitives and extrapolate seen compositions to unseen ones. Existing methods mainly focus on learning a shared embedding space for object-state compositions (Li et al. 2020; Naeem et al. 2021; Nagarajan and Grauman 2018; Khan et al. 2023) or compositional attribute and object classifiers (Purushwalkam et al. 2019; Misra, Gupta, and Hebert 2017; Li et al. 2022; Xu et al. 2022; Yang et al. 2022).

However, the performances of these methods degrade to some extent (Mancini et al. 2021, 2022) as for the open-world setting (OW-CZSL), where there are no priors on the unseen compositions, and the model must consider the whole possible compositions in terms of all objects and states. To deal with such a problem, existing mainstream methods utilize feasibility constraints on the composition embedding (Mancini et al. 2021, 2022) or independently predict simple state and object primitives (Karthik, Mancini, and Akata 2021, 2022). While (Mancini et al. 2021, 2022) rely on different word embedding methods. The straightforward but effective Visual Product method like (Karthik, Mancini, and Akata 2022) predicts the state and object primitives while ignoring the compatibility between two primitives. So external knowledge is introduced to eliminate less feasible compositions, while it is cumbersome to select proper external knowledge for varying datasets.

To address the aforementioned problems, we propose Progressive Cross-primitive Compatibility (ProCC) network to recognize compositions in the open-world setting and a more realistic setting (i.e., partial supervision), aiming at attaining cross-primitive compatibility during easy-hard recognition progress, as shown in Figure 1. Specifically, following the route of the human learning process (Hochstein and Ahissar 2002), we **first** learn to classify objects, which is easier than recognizing states (Saini, Pham, and Shrivastava 2022; Karthik, Mancini, and Akata 2022) because the same state varies greatly conditioned on objects and related contexts, i.e., *ancient castle* / *ancient coin*, and different states are sometimes less feasible composed with the same object, i.e., *old dog* / *ripe dog*. **Then**, with the learned knowledge

*Corresponding author

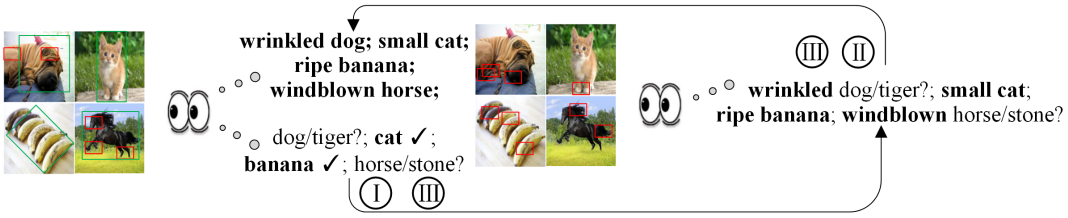

Figure 1: The overall concept of our method. Following the principle of 'forest before trees' (Hochstein and Ahissar 2002), human feedforward hierarchy underlies implicit processing for initial vision at a glance (i.e., green rectangle), and feedback connections add details to explicit vision with scrutiny (i.e., red rectangle). As for composition generalization learning, humans first (Ⅰ) learn to recognize overall objects, then (Ⅱ) gradually identify the scrutiny attribute of objects, i.e., state, and finally (Ⅲ) reasonably compose the object and state primitives. Inspired by this, we aim to progressively recognize the object and state primitives and guide the network to exploit discriminative information conditioned on learned knowledge via the CPC module.

of object primitive, we sequentially classify state primitives conditioned on object features via Cross-Primitive Compatibility (CPC) module, excavating discriminative information. **Finally**, we finetune the whole network conditioned on prior knowledge of two primitives. The ProCC achieves cross-primitive compatibility by adjusting the visual attention to filter out less feasible compositions, without the aid of external knowledge like Word2vec (Mikolov et al. 2013), Glove (Pennington, Socher, and Manning 2014), Conceptnet (Speer, Chin, and Havasi 2017) etc. Also, the progressive training paradigm effectively models the interactions of primitives via conditioned features, especially for pCZSL, where only partial label results in invalid interactions.

In summary, our contributions are four-fold:

1) We propose a novel Progressive Cross-primitive Compatibility (ProCC) network, mimicking the human learning progress of recognizing the state and object compositions without external knowledge.

2) We revisit Visual Product methods and present a Cross-Primitive Compatibility (CPC) module to model the interactions of classifiers to exploit the discriminative visual attention conditioned on each other, guiding the model to generalize to feasible compositions.

3) The progressive training paradigm alleviates the invalid cross-primitive interactions without the aid of cumbersome external knowledge, especially for pCZSL.

4) Comprehensive experimental results on three large-scale datasets for OW-CZSL and pCZSL tasks demonstrate the effectiveness of our proposed approach, which outperforms the state-of-the-art methods[1].

Related Work

Compositional Zero-shot Learning. Compositional Zero-shot Learning (CZSL) aims to recognize the state and object from the images, and even the state-object compositions are not ever seen in the training datasets. Different from typical zero-shot learning (Xian et al. 2019; Huynh and Elhamifar 2020; Li et al. 2021), which aims to utilize attributed vectors or inherent semantic descriptions to recognize unseen instances, The main challenge of CZSL is modeling the relation and affordance of states and objects, generalizing this capability to unseen compositions. Existing methods mainly deal with CZSL in two ways. The first way is inspired by Biederman's Recognition-ByComponents theory(Biederman 1987) and Hoffman's part theory (Hoffman and Richards 1984). For instance, Misra *et al.* (Misra, Gupta, and Hebert 2017) learns a transformation between individual classifiers of states and objects. Other representative methods learn a hierarchical decomposition and composition of the state and object primitives (Yang et al. 2020; Hao, Han, and Wong 2023; Hu and Wang 2023), model objects to be symmetric under attribute transformations (Li et al. 2020), and learn independent prototypical representations of visual primitives then propagated prototype via a compositional graph (Ruis, Burghouts, and Bucur 2021). The second way tries to learn the joint representation of the state-object compositions from given images. Specially, SymNet (Li et al. 2020) enforces symmetries in the representation of objects given their state transformations. Graph network is also employed in (Naeem et al. 2021) to enforce the compositional information transfer from seen to unseen compositions. AoP (Nagarajan and Grauman 2018) regards attribute as the operator and models each state as a linear transformation of objects. CANet (Wang et al. 2023) learns conditional attributes to enhance embedding space. LAP (Khan et al. 2023) exploits the self-attention mechanism to embed related compositions closer and unrelated far away. Differently, causality-based methods (Atzmon et al. 2020; Yang et al. 2022) explore decomposable objects and state representations.

Open-world Compositional Zero-shot Learning. Above methods perform well on the close-world CZSL, while suffering from severe degradation for the open-world setting (Mancini et al. 2021, 2022; Karthik, Mancini, and Akata 2022), where the output space has not imposed any limit. Mancini *et al,* (Mancini et al. 2021) compute feasibility scores (i.e., cosine similarity) between visual features and compositional embeddings to reduce the output space. Then they further inject the feasibility scores both at the loss level and within the graph connections (Mancini et al. 2022). (Karthik, Mancini, and Akata 2022) follows the Visual Product (Misra, Gupta, and Hebert 2017) and predicts state and object primitives independently with non-linear feature extractors. To refine the relation between independent prim-

[1]Codes is in https://github.com/huofushuo/procc and appendix is in https:// arxiv.org/abs/2211.12417

itives, Conceptnet (Speer, Chin, and Havasi 2017) is introduced as the external knowledge. We revisit the Visual Product and achieve cross-primitive compatibility in an easy-hard learning manner, avoiding the external knowledge in (Karthik, Mancini, and Akata 2022) and cumbersome word embeddings in (Mancini et al. 2021, 2022).

Approach

Problem Formulation

Compositional Zero-Shot Learning (CZSL) aims to recognize the composition of two primitives, i.e., an state ($e.g., tiny$) and an object (e.g., dog). Given S and O as two sets of states and objects, spanning *all* classes, we compose a set of possible state-object pairs, i.e., $C = S \times O = \{(s, o) | s \in S, o \in O\}$. Formally, given a training set $D^s = \{(i, c) | i \in I^s, c \in C^s\}$, where I^s is an training image set, and C^s is the corresponding state-object labels. The close world CZSL follows the generalized ZSL (Xian et al. 2019) that the test sample comes from either seen (C^s) or unseen (C^u) composition ($C^s \cup C^u$). For the **Open-World CZSL (OW-CZSL) setting** (Mancini et al. 2021), there assumes no prior on the set of testing compositions. It means the model must consider the full compositional space (C), which is much larger than $C^s \cup C^u$. Consequently, the unseen compositions are $C^u_{ow} = C \backslash C^s$. OW-CZSL introduces a more practical setting while bringing more challenging problems: 1) It is hard to generalize from small seen compositions to large unseen compositions. 2) There are a large number of less feasible compositions in the full composition space (C), confusing the prediction models. Recently, (Karthik, Mancini, and Akata 2022) proposes a new practical setting, i.e., only training with one of the state and object annotations, named **partial-supervision CZSL (pCZSL)**. Formally, for the training set C^s, The relation of the partial label of state and object primitives can be formulated as: $\{(s, u)\} \cup \{(u, o)\} = C^s$, where u indicates unlabeled primitives. Consequently, the test set in pCZSL has the full output composition space (C) like OW-CZSL, while the training set in pCZSL does not have the composition knowledge about any state-object pairs. Therefore, the joint training strategy may fail due to lacking the explicit supervision to learn how states interact with objects and vice-versa.

Progressive Cross-primitive Compatibility (ProCC)

Most CZSL methods (Atzmon et al. 2020; Li et al. 2020; Nagarajan and Grauman 2018; Purushwalkam et al. 2019; Saini, Pham, and Shrivastava 2022; Mancini et al. 2021, 2022; Naeem et al. 2021) explicitly modulate the interactions of states and objects to improve the generalization ability. However, it is less effective for OW-CZSL and pCZSL due to large output space and missing labels. Some methods (Karthik, Mancini, and Akata 2021, 2022) follow the Visual Product (Misra, Gupta, and Hebert 2017) that independently predict the state and object primitives, disregarding compositional nature. Following the route of (Karthik, Mancini, and Akata 2021, 2022; Misra, Gupta, and Hebert 2017), we propose Progressive Cross-primitive Compatibility (ProCC) network while achieving cross-primitive compatibility. Also, like the human learning process (Hochstein and Ahissar 2002), ProCC trains the network in an easy-hard manner, which dynamically models interactions between state and object primitives, alleviating the negative influence of no explicit supervision on both states and objects in pCZSL. Figure 2 shows the framework of the proposed approach. In the following subsections, we revisit the Visual Product and introduce a cross-primitive compatibility module and progressive learning strategy.

Revisit Visual Product. Generally, given an image i, CZSL wants to model the joint probability distribution $p(s_i, o_i|i)$. The visual product simplifies this as follows:

$$p(s_i, o_i|i) \approx p(s_i|i) \times p(o_i|i) \tag{1}$$

In this way, Visual Product treats the states and objects *independently* only from the visual cues, without side information (i.e., word embeddings). Concretely, input image i is firstly encoded to obtain the feature z as: $z = \omega(i)$. Then the object (i.e., $\varphi_o \langle z, o \rangle$) and state (i.e., $\varphi_s \langle z, s \rangle$) classifiers assign z to the vectors in the probability simplex o and s, spanning *all* object and state classes. Visual Product minimizes the cross-entropy loss of seen compositions ($D^s = \{I^s, C^s\}$) for both object and state predictions:

$$\ell_{vp} = \ell_{obj}(i, o_i) + \ell_{state}(i, s_i) \tag{2}$$

$$\ell_{obj} = \min_{\varphi_o} \sum \ell_{ce}(\varphi_o \langle \omega(i), o \rangle, o_i) \tag{3}$$

$$\ell_{state} = \min_{\varphi_s} \sum \ell_{ce}(\varphi_s \langle \omega(i), s \rangle, s_i) \tag{4}$$

where $(i, (s_i, o_i)) \in D^s$. Thus, the prediction function is:

$$f(i) = \arg \max_{(s,o) \in C} \varphi_s \langle \omega(i), s \rangle \times \varphi_o \langle \omega(i), o \rangle \tag{5}$$

where C represents the full state-object composition pairs in OW-CZSL. As the search space is huge, Visual Product is more effective than previous methods, which aim to produce discriminative state-object embeddings (Karthik, Mancini, and Akata 2021, 2022). Recently, (Karthik, Mancini, and Akata 2021, 2022) expanded the visual product and equipped the classifiers with multi-layer perceptrons (MLP) to excavate discriminative features. Also, external knowledge (Speer, Chin, and Havasi 2017) is employed in (Karthik, Mancini, and Akata 2022) to estimate the feasibility scores of compositions. *Here*, we explicitly model the composition interactions via Cross-Primitive Compatibility (CPC) module during the training procedure, without external knowledge. Also, considering the pCZSL setting and better modulating the primitive compatibility, the progressive learning strategy, following the human learning process (Hochstein and Ahissar 2002), is proposed to facilitate cross-primitive compatibility in an easy-hard manner.

Cross-primitive Compatibility Module. Visual Product methods independently predict compositions via Equation 1, which ignores the fact that the feasibility of state-object compositions is heavily *conditioned* on each other. A more practical compositional probability can be modeled as:

$$p(s_i, o_i|i) \approx p(s_i|i, f_o(i))) \times p(o_i|i, f_s(i))) \tag{6}$$

where $f_o(i)$ and $f_s(i)$ are intermediate features of the object and state primitives. It is non-trivial to directly model

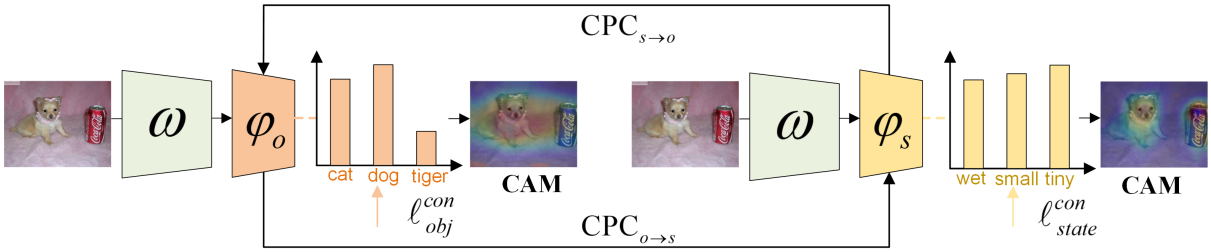

Figure 2: The framework of ProCC. Features from the encoder (ω) are respectively fed to the object and state (φ_o and φ_s) classifiers, where the Cross-Primitive Compatibility (CPC) aims to model the cross-primitive interactions. Progressive learning strategy is proposed to gradually modulate primitive compatibility, especially for pCZSL. For detailed training procedure, please refers to Appendix: Algorithm 1. Class Activation Maps (CAM) of input samples are illustrated to show visual attention.

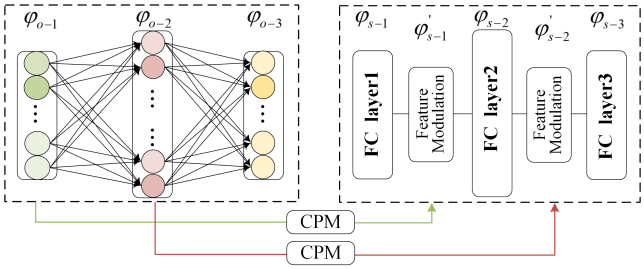

Figure 3: The detailed framework of the object-state Cross-Primitive Compatibility (CPC$_{o\to s}$). Features from the object classifier (φ_{o-1} and φ_{o-2}) are encoded by learnable Cross-Primitive Memory (CPM) units. Then respectively interact with state features (φ_{s-1} and φ_{s-2}) to achieve compatibility of state features conditioned on objects.

the relationship between objects and states due to the diverse semantic entanglement and a large number of possible compositions. We integrate the feasibility reasoning into the trainable Cross-Primitive Compatibility (CPC) module, which facilitates interactions between two classifiers to explore informative visual attention conditioned on feature representations of each primitive. Specifically, The features extracted by the encoder (ω) are fed to primitive classifiers (i.e., φ_o and φ_s). The primitive classifiers *follow* the Visual Product methods (Karthik, Mancini, and Akata 2021, 2022) that consist of multi-layer perceptron (MLP), specifically three-layer MLP, for classifications. As shown in Figure 2 and Equation 6, the network is *symmetric* and we take the object-state CPC (CPC$_{o\to s}$) module for example, as shown in Figure 3, intermediate features from φ_{o-1} and φ_{o-2} are fed to φ_s to interact with state features. However, direct modulation state features will induce information degradation because of the huge task diversity. We propose learnable Cross-Primitive Memory (CPM) units for soft interactions. Specifically, the learnable CPM unit introduces conditioned information to modulate corresponding features along with the residual connection, which is formulated as follows:

$$\varphi^m_{o-l} = \sigma\left(\mathrm{Conv}^k_{1\mathrm{d}}\left(\varphi_{o-l}\right)\right), l \in (1,2) \tag{7}$$

$$\varphi'_{s-l} = \varphi_{s-l} \times \varphi^m_{o-l} + \varphi_{s-l}, l \in (1,2) \tag{8}$$

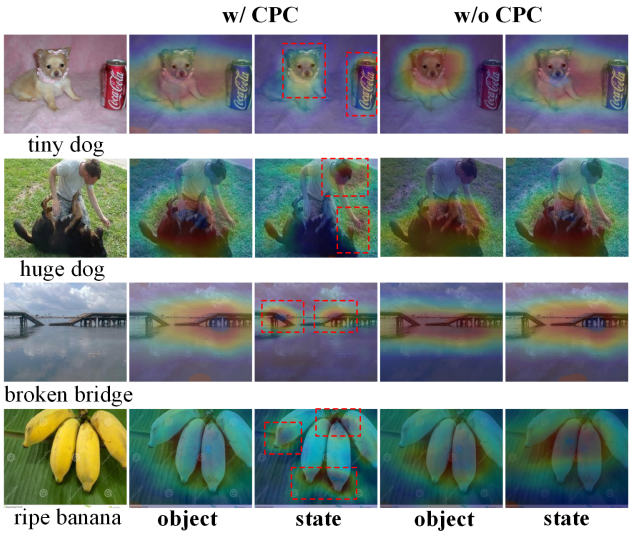

Figure 4: Visualizations of class activation maps of ProCC with and without CPC modules on the testing dataset of MIT-States. The discriminative regions are marked with red rectangles. More visualizations are in Appendix 2.

where $\mathrm{Conv}^k_{1\mathrm{d}}$ and σ represent the 1d convolution layer and softmax activation function. Kernel size (k) is equal to 1/10 feature dimension to efficiently capture the long-range dependency. For the hyper-parameter analysis of CPC, please refers to **Appendix 3**. Then the enhanced state features are fed to the next layer of φ_s as:

$$\varphi_{s-(l+1)} = f_{s-l}(W^T_{s-l}\varphi'_{s-l} + b_{s-l}), l \in (1,2), \tag{9}$$

where W and b are weights and biases of MLP. Accordingly, the conditioned cross-primitive interactions are injected into each other, reducing less feasible primitive predictions. Therefore, Equations 3 and 4 can be re-write as:

$$\ell^{con}_{obj} = \min_{\varphi_o,\varphi_{o\to s}} \sum \ell_{ce}(\varphi_o\left\langle z|\varphi_{s\to o}(\varphi_s(z)), o\right\rangle, o_i) \tag{10}$$

$$\ell^{con}_{state} = \min_{\varphi_s,\varphi_{s\to o}} \sum \ell_{ce}(\varphi_s\left\langle z|\varphi_{o\to s}(\varphi_o(z)), s\right\rangle, s_i) \tag{11}$$

where $z = \omega(i), (i,(s_i,o_i)) \in D^s$, and $\ell^{con}_{vp} = \ell^{con}_{obj} + \ell^{con}_{state}$

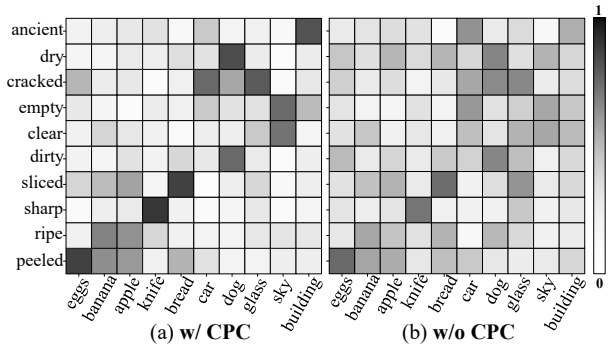

Figure 5: Confusion matrices about prediction probabilities of states conditioned on objects (w/ CPC) or not (w/o CPC).

Visual Explanation. To further illustrate and explain the effect of the CPC module, we visualize the attention learned from the classifier via Class Activation Map (CAM) (Zhou et al. 2016) in Figure 4. The standard CAM is formulated as:

$$\mathrm{CAM}_c(x, y) = \sum_k \omega_k^c f_k(x, y) \tag{12}$$

where $\mathrm{CAM_c}$ means the class activation map that leads to the classification of an image to class c. $f_k(x, y)$ and ω_k^c stand for the activation of unit k in the last layer at spatial location (x, y) and the weight corresponding to class c for unit k. Here, ω_k^c is the final layer of the MLP (i.e., φ_{o-3} and φ_{s-3}), which has been modulated by the CPC modules. Figure 4 shows some visualization examples with (w/) and without (w/o) CPC module. As the encoder (ω) is pre-trained for the object classification task, most CAMs for the object classifier can locate and recognize the proper attention regions. However, the CAMs for the state classifier vary greatly as state primitives are conditioned on the object primitive and related contexts. For the $tiny\ dog$ and $huge\ dog$ compositions, the CPC module drives the model to focus on the discriminative regions that a dog with a small head compared with other objects tends to classify to the $tiny$ otherwise classify to $huge$. For more abstract compositions, $broken\ bridge$ and $ripe\ banana$ compositions, the state primitives heavily depend on the object primitives otherwise may induce less feasibility compositions. The state of $broken$ is mainly reflected in the curvatures of the bridge and the $ripe$ primitive of the banana displays the black spots on the surface. Overall, the CPC module enables the efficient adjustment of visual attention conditioned on mutual relations. Moreover, Figure 5 illustrates the confusion matrices about state and object primitives. Concretely, we select ten typical state and object primitives in the MIT-States (Isola, Lim, and Adelson 2015) dataset. Prediction probabilities of states are accumulated then normalized with and without CPC module to formulate the confusion matrices (Confusion matrices of objects prediction probabilities are in **Appendix 2**). We can learn that the CPC module facilitates reasoning compatible compositions with high confidence.

Progressive Learning Strategy. However, jointly training the state and object classifiers may induce two issues: **(1)** When it comes to the more practical setting, partial supervision Compositional Zero-Shot Learning (pCZSL), where only the partial label, not both, is available (Karthik, Mancini, and Akata 2022). The missing label makes the joint training strategy invalid to model the interactions between the state and object primitives. A naive way of learning from such partial supervision is to update the parameters of the state and object classifier only based on the available labels, which lacks the interaction information across primitives via the CPC module. Recent method (Karthik, Mancini, and Akata 2022) estimates the missing labels via pseudo-labeling (Lee 2013) as well as utilizes the external knowledge (Speer, Chin, and Havasi 2017). The challenge of missing labels also exists in the standard Multi-Task Learning (MTL) that the traditional updating rule will give inferior results due to the missing annotations (Vandenhende et al. 2022; Kim et al. 2018; Nekrasov et al. 2019; Li, Liu, and Bilen 2022). Some typical solutions propose hard knowledge distillation (Kim et al. 2018), alternative optimization strategy (Nekrasov et al. 2019), and learning in the joint pairwise task spaces (Li, Liu, and Bilen 2022). However, compared with the MTL task, the missing label issue matters more to the CZSL task, as the object and state primitives are heavily tangled. **(2)** Also, jointly training results in suboptimal interactions as the diverse difficulty of object and state predictions. Concretely, classifying states is *more* challenging than objects (Saini, Pham, and Shrivastava 2022; Karthik, Mancini, and Akata 2022). Therefore, joint training inevitably induces noisy conditioned information, which hinders to reason cross-primitive compatibility. Quantitative analysis is shown in **Appendix 4**.

To enable the full interaction of state and object primitives, we propose a progressive learning strategy, mimicking the easy-hard learning process shown in Figure 1. Concretely, with the features from the encoder (ω), we first train the object classifier φ_o with given labels (Equation 3), to obtain object features (φ_{o-l}, $x \in (1, 2)$). Then we sequentially train the state classifier φ_s and $\mathrm{CPC}_{o\to s}$ ($\varphi_{o\to s}$) conditioned on pre-trained object features (φ_{o-l}) (Equation 11), to interact to adjust the visual attention. Finally, we fine-tune the state and object classifiers (φ_s and φ_o) as well as CPC modules ($\varphi_{o\to s}$ and $\varphi_{s\to o}$) conditioned on the well-trained features (Equations 10 and 11). We utilize this training protocol both in the OW-CZSL and pCZSL settings. During the easy-hard recognition progress, our method alleviates invalid interactions of cross primitives, especially in the pCZSL setting, without external knowledge. For detailed training procedure, please refers to **Algorithm 1**.

Experiments

Datasets and Evaluation Metrics. We conduct experiments on three widely-use datasets including UT-Zappos (Yu and Grauman 2014), MIT-States (Isola, Lim, and Adelson 2015), and C-GQA (Misra, Gupta, and Hebert 2017). Details of three datasets are listed in **Appendix 1**. For the **OW-CZSL**, we follow the splits of (Mancini et al. 2021, 2022; Karthik, Mancini, and Akata 2022) and evaluate based on the generalized settings, where the test samples are from both seen and unseen compositions. Considering the performance of the model with different bias factors for the unseen com-

Algorithm 1: Training procedure of ProCC.

Input: Training data $D^s = \{(i, c)|i \in I^s, c \in C^s\}$, pre-trained ω, learning rate λ_1, λ_2, λ_3
Output: Optimal φ_o, φ_s, CPC: $\varphi_{o\to s}$, $\varphi_{s\to o}$

Initialize: φ_o, φ_s, $\varphi_{o\to s}$, $\varphi_{s\to o}$;
Stage 1: *// train* φ_o
while *not converged* **do**
  Sample a batch from D^s as images $(i_k)_{k=1}^n$ with their object labels $(o_k)_{k=1}^n$;
  for *samples in the batch* **do**
    Compute ℓ_{obj} via Equation 3.;
    Update $\varphi_o \leftarrow \varphi_o - \lambda_1 \nabla_{\varphi_o} \ell_{obj}$
Stage 2: *// train* φ_s *and* $\varphi_{o\to s}$
while *not converged* **do**
  Sample a batch from D^s as images $(i_k)_{k=1}^n$ with their state labels $(s_k)_{k=1}^n$;
  for *samples in the batch* **do**
    Compute ℓ_{state}^{con} via Equation 11.;
    Update $\varphi_{s\cup o\to s} \leftarrow \varphi_{s\cup o\to s} - \lambda_2 \nabla_{\varphi_{s\cup o\to s}} \ell_{state}^{con}$
Stage 3: *// finetune* φ_o, φ_s, $\varphi_{o\to s}$, *and* $\varphi_{s\to o}$
while *not converged* **do**
  Sample a batch from D^s as images $(i_k)_{k=1}^n$ with their object and state labels $(o_k, s_k)_{k=1}^n$;
  for *samples in the batch* **do**
    Compute ℓ_{vp}^{con} via Equations 10 and 11.;
    Update $\varphi_{total} \leftarrow \varphi_{total} - \lambda_3 \nabla_{\varphi_{total}} \ell_{vp}^{con}$

positions, we vary the bias on the seen composition (C^s) during the test phase and report the performance as best seen (S), best unseen (U), best harmonic mean (HM), and the Area Under the Curve (AUC). For the **pCZSL**, following (Karthik, Mancini, and Akata 2022), we remove the label and calculate the metrics on the full output composition space (C). As we can not access the full-labeled seen compositions (C^s), we do not subtract any bias on C^s. Therefore, we use the seen (S), unseen (U), and HM metrics.

Baselines. For **OW-CZSL**, we compare ProCC with other OW-CZSL methods, including CompCos (Mancini et al. 2021), KGSP (Karthik, Mancini, and Akata 2022), and Co-CGE (Mancini et al. 2022). CZSL methods are also compared, including LE+ (Misra, Gupta, and Hebert 2017), AoP (Nagarajan and Grauman 2018), TMN (Purushwalkam et al. 2019), SymNet (Li et al. 2020), CGE (Naeem et al. 2021), and CANet (Wang et al. 2023). For **pCZSL**, ProCC is compared with KGSP (Karthik, Mancini, and Akata 2022) as well as standard (OW-)CZSL methods like CGE (Naeem et al. 2021), CompCos (Mancini et al. 2021), and Co-CGE (Mancini et al. 2022), with the same partial label protocol.

Implementation Details. Following the standard protocols in the CZSL, we utilize the pre-trained ResNet-18 (He et al. 2016) as the feature encoder (ω) to extract 512-dimensional feature vectors and learn classifiers on top of these features. Following (Naeem et al. 2021; Karthik, Mancini, and Akata 2022), each classifier is composed of Multi-Layer Perceptrons (MLP) with three layers with dimensions 768, 512, and the number of output classes, respectively, and comprise Layer Normalization(Lei Ba, Kiros, and Hinton 2016) and Dropout(Srivastava et al. 2014). To be consistent with other methods, we randomly augment input images with random crop and horizontal flip. We use PyTorch to implement our network and optimize it with Adam (Kingma and Ba 2015) with default settings. The batch size is 256, and the learning rate is 5.0×10^{-5} for the first two stages and 1.0×10^{-5} for the third stage. For the UT-Zappos, MIT-States, and C-GQA datasets, the total training time is approximately 1, 3, and 5 hours for 30/60/20, 40/80/30, and 50/100/25 epochs for three stages, respectively, with the early stop strategy.

Open-World CZSL (OW-CZSL) Results

The results of OW-CZSL setting are illustrated in Table 1. Generally, closed-world CZSL methods achieve inferior performance, especially in two large datasets (i.e., C-GQA and MIT-States), due to the large cardinality of the output space. ProCC outperforms previous methods on almost all metrics in terms of three datasets. Concretely, as for the most challenging dataset, i.e., C-GQA, the proposed method exceeds the previous SOTA methods, especially for best harmonic (HM) metrics (3.4→3.8: ↑12%), which means that ProCC has the better ability to recognize both the seen and unseen compositions. Also, in the validation sub-dataset, Our method suppresses the best baseline (i.e., KGSP) by a large margin in two overall evaluation indexes (i.e., HM: 13.2→16.1: ↑22%; AUC: 2.9→4.0: ↑38%). As for the MIT-States dataset, our method also has comparative results. Notably, we achieve the best performance on the U metric, which validates the generalization ability of ProCC. For UT-Zappos, it is specially designed for shoes and is relatively simpler than others. ProCC consistently outperforms others, i.e., S: 59.3→62.2; U: 47.2→48.0; HM: 39.1→39.9; AUC: 22.9→23.6. Remarkably, previous methods typically utilize word embeddings to encode the word expression, which already contains semantic knowledge of similar objects and attributes for composition learning (Saini, Pham, and Shrivastava 2022). Recent Visual Product based method (Karthik, Mancini, and Akata 2022) employs more complex classifiers (with hidden layers of 768 and 1024) than ours as well as uses external knowledge to eliminate the less feasibility compositions. We predict the state and object primitives with more lightweight classifiers and explicitly model the cross-primitive interactions to learn the relationship between primitives without external knowledge.

Partial-supervision CZSL (pCZSL) Results

As for the more challenging setting, pCZSL, the challenges come from not only the huge output composition space but also the *missing* labels. As we can learn from Table 2, our method achieves SOTA performances compared with previous CZSL, OW-CZSL, and pCZSL methods. Concretely, for the largest dataset, C-GQA, the performance of SOTAs on pCZSL severely degrades compared with OW-CZSL, even for KGSP, which is equipped with the pseudo label and external knowledge. Our method consistently exceeds

Method	C-GQA						MIT-States						UT-Zappos					
	Val		Test				Val		Test				Val		Test			
	HM	AUC	S	U	HM	AUC	HM	AUC	S	U	HM	AUC	HM	AUC	S	U	HM	AUC
TMN	NA	NA	NA	NA	NA	NA	2.1	0.2	12.6	0.9	1.2	0.1	21.2	9.2	55.9	18.1	21.7	8.4
AoP	NA	NA	NA	NA	NA	NA	3.2	0.3	16.6	5.7	4.7	0.7	23.4	10.1	50.9	34.2	29.4	13.7
LE+	9.3	1.8	19.2	0.7	1.0	0.08	5.3	0.5	14.2	2.5	2.7	0.3	26.6	14.3	60.4	36.5	30.5	16.3
VisProd	10.5	2.0	24.8	1.7	2.8	0.33	7.2	1.0	20.9	5.8	5.6	0.7	28.8	15.4	54.6	42.8	36.9	19.7
SymNet	12.3	2.5	26.7	2.2	3.3	0.43	8.0	1.2	21.4	7.0	5.8	0.8	32.5	16.7	53.3	44.6	34.5	18.5
CGE	12.8	2.8	28.3	1.3	2.2	0.30	8.3	1.8	**29.6**	4.0	4.9	0.7	34.5	18.9	58.8	46.5	38.0	21.5
CompCos	12.0	2.4	28.4	1.8	2.8	0.39	8.4	1.5	25.4	10.0	8.9	1.6	32.5	18.1	59.3	46.8	36.9	21.3
Co-CGE	12.3	2.7	28.7	1.6	2.6	0.37	8.4	**2.1**	26.4	10.4	**10.1**	**2.0**	34.8	19.2	60.1	44.3	38.1	21.3
KGSP	13.2	2.9	26.6	2.1	3.4	0.44	7.9	1.4	23.4	7.0	6.7	1.0	33.2	19.8	58.0	47.2	39.1	22.9
CANet	14.3	2.8	27.3	1.9	3.2	0.39	8.3	1.7	25.3	6.7	6.6	1.2	35.1	19.8	58.7	46.0	38.7	22.1
Ours	**16.1**	**4.0**	**29.0**	**2.6**	**3.8**	**0.54**	**8.6**	1.9	27.6	**10.6**	7.8	1.6	**36.5**	**22.4**	**62.2**	**48.0**	**39.9**	**23.6**

Table 1: Quantitative comparisons in the OW-CZSL setting. We report the best seen (S), best unseen (U) accuracy, HM, AUC on the test and validation sub-datasets. The best and second-best results are bold and underlined.

Method	C-GQA						MIT-States						UT-Zappos					
	Val			Test			Val			Test			Val			Test		
	S	U	HM	S	U	HM	S	U	HM	S	U	HM	S	U	HM	S	U	HM
CGE	19.2	2.9	5.6	17.4	0.4	0.9	10.0	2.8	4.3	**19.6**	1.3	2.4	46.5	3.5	6.6	50.3	3.4	5.0
CompCos	18.2	3.0	5.2	24.3	0.4	0.7	11.1	2.9	4.6	10.8	2.0	3.6	50.2	3.9	7.3	52.4	4.1	7.6
Co-CGE	19.8	3.9	6.4	22.1	0.6	1.2	14.8	3.3	5.3	13.1	2.3	4.0	47.2	6.1	10.8	52.6	5.4	9.9
KGSP	20.1	4.8	8.3	22.3	0.9	1.7	15.7	3.2	5.3	13.5	2.6	4.4	49.4	5.9	9.7	53.8	6.9	12.3
Ours	**21.6**	**5.4**	**8.7**	**24.1**	**1.1**	**2.0**	**16.3**	**3.5**	**5.8**	14.1	**2.9**	**4.8**	**51.0**	**7.1**	**12.5**	**55.1**	**8.1**	**14.1**

Table 2: Quantitative comparisons in the pCZSL setting. We report the seen (S), unseen (U) accuracy, and best harmonic mean (HM) on the test and validation sub-datasets. The best and second-best results are bold and underlined.

Method	OW-CZSL				pCZSL					
	C-GQA		MIT-States		C-GQA			MIT-States		
	HM	AUC	HM	AUC	S	U	HM	S	U	HM
w/o CPC	3.3	0.40	6.2	0.8	17.4	0.5	1.0	11.6	2.2	3.7
w/o CPI	3.4	0.41	6.1	0.9	17.7	0.5	1.0	12.0	2.1	3.6
w/o CPM	3.5	0.48	6.6	1.0	18.9	0.7	1.4	12.2	2.5	4.1
w/o P-L	3.7	0.50	7.6	1.5	22.4	0.8	1.6	12.5	2.5	4.1
w/ Ex-1&2	3.6	0.48	7.8	1.5	22.6	1.0	1.9	13.2	2.7	4.4
w/o Stage3	3.5	0.47	7.4	1.4	23.2	1.1	2.0	13.6	2.8	4.6
w/ 4 Stages	3.6	0.50	7.6	1.4	23.7	1.0	1.9	13.8	2.8	4.7
w/ 5 Stages	3.7	0.53	7.7	1.4	23.9	1.1	2.1	13.6	2.9	4.8
w/ 6 Stages	3.8	0.56	7.7	1.6	24.0	1.1	2.1	13.8	2.8	4.7
Ours	3.8	0.54	7.8	1.6	24.1	1.1	2.0	14.1	2.9	4.8

Table 3: Ablation studies for both OW-CZSL and pCZSL.

them both on validation and testing datasets. For the MIT-States dataset, our method surpasses the second-best method by a large margin in HM metric (i.e., val: 5.3→5.8:↑9%; test: 4.4→4.8:↑9%). For the simplest dataset, UT-Zappos, our method also has the best performance. Note that we do not use any external knowledge like Word2vec, Glove, Conceptnet, and other semi-supervised learning techniques (Lee 2013; Grandvalet and Bengio 2004) for the missing annotations. The superior performance indicates even with partial labels of object and state primitives, our progressive learning strategy can also model the interactions of cross primitives with the pre-trained classifiers.

Ablation Study

We analyze two important components: Cross-Primitive Compatibility (CPC) module and the progressive learning strategy. We adopt the same implementation strategy and conduct the OW-CZSL and pCZSL experiments on the two largest datasets, i.e., C-GQA and MIT-States.

Effect of the Cross-Primitive Compatibility Module. In Table 3, ① without the CPC module (w/o CPC), the performance is severely degraded both on the OW-CZSL and pCZSL settings. Because lacking the interaction between cross primitives makes the network degenerate to previous Visual Product baselines (Karthik, Mancini, and Akata 2021, 2022). Meanwhile, KGSP utilizes the external knowledge and surpasses the ablation configuration, especially in pCZSL setting. ② Moreover, to further evaluate the conditional modulation, we employ channel attention (Hu, Shen, and Sun 2018; Wang et al. 2020) on the same primitive classifiers without cross-primitive interaction (w/o CPI). ③ Also, we ablate the learnable cross-primitive memory (w/o CPM) and directly modulate other primitives with learned features. Results indicate that exploring internal primitives brings marginal improvement for composition learning as classifiers have extracted enough internal information, and modulating primitives via hard masks also gives sub-optimal results. Note that the CPC is extremely lightweight with two trainable 1d convolution layers. ④ Besides, more ablations about architectures of CPC and classifiers are in **Appendix 3**. Generally, the CPC module greatly improves the perfor-

mance with negligible computation burden also without external information, which is practical for real-world scenes.

Effect of the Progressive Learning Strategy. Another important aspect of the ProCC is the progressive learning strategy. From Table 3, ① we can learn that with the traditional end-end training strategy (w/o P-L), the performance of ProCC degrades to some extent, *especially* in the pCZSL setting (i.e., HM: 2.0→1.6 (C-GQA) and 4.8→4.1 (MIT-States)). As jointly training the whole network under the pCZSL setting does not explicitly learn the relationship between state and object primitives, which is the critical issue in the CZSL task. While for the OW-CZSL setting, joint training induces some noisy conditioned information, due to the diverse difficulty of classifying object and state primitives. Also, we exchange the training sequence (i.e., Stage $2 \rightarrow 1 \rightarrow 3$) (w/ Ex-1&2) and ablate the fine-tuning stage (w/o Stage 3). ② For the configuration of w/ Ex-1&2, the performance of ProCC degrades on both settings. Due to the challenge of classifying state primitives (Saini, Pham, and Shrivastava 2022; Karthik, Mancini, and Akata 2022), modulation object features conditioned on noisy state features results in invalid interactions. ③ For the configuration of w/o Stage 3, where only $\text{CPC}_{o\rightarrow s}$ works, the performance degrades to some extent. We have two observations: $\text{CPC}_{o\rightarrow s}$ brings *more* improvements than $\text{CPC}_{s\rightarrow o}$; $\text{CPC}_{s\rightarrow o}$ and fine-tuning based on well-trained features also matter for the cross-primitive compatibility and global optimum. ④ Moreover, following the same training protocol, we train the network for more stages, i.e., with extra Stage 1 (w/ 4 Stages), extra Stage 1 and 2 (w/ 5 Stages), and extra Stage 1, 2, and 3 (w/ 6 Stages). We see that more training stages can not bring much accuracy improvement, as the model has converged after Stage 3.

Conclusion

In this paper, we propose a method named Progressive Cross-primitive Compatibility (ProCC) network for both OW-CZSL and pCZSL tasks. The simple but effective Cross-Primitive Compatibility module drives the network learning to predict feasible object and state primitives conditioned on mutual relations. Also, the progressive learning strategy significantly eliminates the invalid cross-primitive interactions in pCZSL and noisy conditioned information, in an easy-hard learning manner. Comprehensive experiments on OW-CSZL and pCZSL settings illustrate superior performance compared with other state-of-the-art methods.

Acknowledgments

This research was partially supported by Project PolyU15222621 and PolyU15225023, the Key-Area Research and Development Program of Guangdong Province (No. 2021B0101400003), Hong Kong RGC Research Impact Fund (No. R5060-19, No. R5034-18), Areas of Excellence Scheme (AoE/E-601/22-R), General Research Fund (No. 152203/20E, 152244/21E, 152169/22E, 152228/23E), Shenzhen Science and Technology Innovation Commission (JCYJ20200109142008673), Hong Kong RGC General Research Fund (No. 152211/23E), the National Natural Science Foundation of China (No. 62102327), and PolyU Internal Fund (No. P0043932), and the National Natural Science Foundation of China under grants 62302184.

References

Atzmon, Y.; Berant, J.; Kezami, V.; Globerson, A.; and Chechik, G. 2016. Learning to generalize to new compositions in image understanding. *arXiv e-prints*, arXiv:1608.07639.

Atzmon, Y.; Kreuk, F.; Shalit, U.; and Chechik, G. 2020. A causal view of compositional zero-shot recognition. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *NeurIPS*, volume 33, 1462–1473.

Biederman, I. 1987. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2).

Grandvalet, Y.; and Bengio, Y. 2004. Semi-supervised Learning by Entropy Minimization. In Saul, L.; Weiss, Y.; and Bottou, L., eds., *NeurIPS*, volume 17.

Hao, S.; Han, K.; and Wong, K.-Y. K. 2023. Learning Attention As Disentangler for Compositional Zero-Shot Learning. In *CVPR*, 15315–15324.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*.

Hochstein, S.; and Ahissar, M. 2002. View from the Top: Hierarchies and Reverse Hierarchies in the Visual System. *Neuron*, 36(5): 791–804.

Hoffman, D.; and Richards, W. 1984. Parts of recognition. *Cognition*, 18(1): 65–96.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-Excitation Networks. In *CVPR*.

Hu, X.; and Wang, Z. 2023. Leveraging Sub-class Discimination for Compositional Zero-Shot Learning. *AAAI*, 890–898.

Huynh, D.; and Elhamifar, E. 2020. Fine-Grained Generalized Zero-Shot Learning via Dense Attribute-Based Attention. In *CVPR*.

Isola, P.; Lim, J. J.; and Adelson, E. H. 2015. Discovering States and Transformations in Image Collections. In *CVPR*.

Karthik, S.; Mancini, M.; and Akata, Z. 2021. Revisiting Visual Product for Compositional Zero-Shot Learning. In *NeurIPS*.

Karthik, S.; Mancini, M.; and Akata, Z. 2022. KG-SP: Knowledge Guided Simple Primitives for Open World Compositional Zero-Shot Learning. In *CVPR*, 9336–9345.

Khan, M. G. Z. A.; Naeem, M. F.; Van Gool, L.; Pagani, A.; Stricker, D.; and Afzal, M. Z. 2023. Learning Attention Propagation for Compositional Zero-Shot Learning. In *WACV*, 3828–3837.

Kim, D.-J.; Choi, J.; Oh, T.-H.; Yoon, Y.; and Kweon, I. S. 2018. Disjoint Multi-task Learning Between Heterogeneous Human-Centric Tasks. In *WACV*, 1699–1708.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization.

Lee, D.-H. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop*.

Lei Ba, J.; Kiros, J. R.; and Hinton, G. E. 2016. Layer Normalization. *arXiv e-prints*, arXiv:1607.06450.

Li, W.-H.; Liu, X.; and Bilen, H. 2022. Learning Multiple Dense Prediction Tasks From Partially Annotated Data. In *CVPR*, 18879–18889.

Li, X.; Xu, Z.; Wei, K.; and Deng, C. 2021. Generalized Zero-Shot Learning via Disentangled Representation. *AAAI*, 35(3): 1966–1974.

Li, X.; Yang, X.; Wei, K.; Deng, C.; and Yang, M. 2022. Siamese Contrastive Embedding Network for Compositional Zero-Shot Learning. In *CVPR*, 9326–9335.

Li, Y.-L.; Xu, Y.; Mao, X.; and Lu, C. 2020. Symmetry and Group in Attribute-Object Compositions. In *CVPR*.

Mancini, M.; Naeem, M. F.; Xian, Y.; and Akata, Z. 2021. Open World Compositional Zero-Shot Learning. In *CVPR*, 5222–5230.

Mancini, M.; Naeem, M. F.; Xian, Y.; and Akata, Z. 2022. Learning Graph Embeddings for Open World Compositional Zero-Shot Learning. *IEEE TPAMI*, 1–1.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *NeurIPS*, volume 26.

Misra, I.; Gupta, A.; and Hebert, M. 2017. From Red Wine to Red Tomato: Composition With Context. In *CVPR*.

Naeem, M. F.; Xian, Y.; Tombari, F.; and Akata, Z. 2021. Learning Graph Embeddings for Compositional Zero-Shot Learning. In *CVPR*, 953–962.

Nagarajan, T.; and Grauman, K. 2018. Attributes as Operators: Factorizing Unseen Attribute-Object Compositions. In *ECCV*.

Nekrasov, V.; Dharmasiri, T.; Spek, A.; Drummond, T.; Shen, C.; and Reid, I. 2019. Real-Time Joint Semantic Segmentation and Depth Estimation Using Asymmetric Annotations. In *ICRA*, 7101–7107.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Purushwalkam, S.; Nickel, M.; Gupta, A.; and Ranzato, M. 2019. Task-Driven Modular Networks for Zero-Shot Compositional Learning. In *ICCV*.

Ruis, F.; Burghouts, G.; and Bucur, D. 2021. Independent Prototype Propagation for Zero-Shot Compositionality. In *NeurIPS*, volume 34.

Saini, N.; Pham, K.; and Shrivastava, A. 2022. Disentangling Visual Embeddings for Attributes and Objects. In *CVPR*, 13658–13667.

Speer, R.; Chin, J.; and Havasi, C. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *AAAI*, 4444–4451.

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; and Salakhutdinov, R. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *JMLR*, 15: 1929–1958.

Vandenhende, S.; Georgoulis, S.; Van Gansbeke, W.; Proesmans, M.; Dai, D.; and Van Gool, L. 2022. Multi-Task Learning for Dense Prediction Tasks: A Survey. *IEEE TPAMI*, 44(7): 3614–3633.

Wang, Q.; Liu, L.; Jing, C.; Chen, H.; Liang, G.; Wang, P.; and Shen, C. 2023. Learning Conditional Attributes for Compositional Zero-Shot Learning. In *CVPR*, 11197–11206.

Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; and Hu, Q. 2020. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In *CVPR*, 11531–11539.

Xian, Y.; Lampert, C. H.; Schiele, B.; and Akata, Z. 2019. Zero-Shot Learning—A Comprehensive Evaluation of the Good, the Bad and the Ugly. *IEEE TPAMI*, 41(9): 2251–2265.

Xu, Z.; Wang, G.; Wong, Y.; and Kankanhalli, M. S. 2022. Relation-Aware Compositional Zero-Shot Learning for Attribute-Object Pair Recognition. *IEEE TMM*, 24: 3652–3664.

Yang, M.; Deng, C.; Yan, J.; Liu, X.; and Tao, D. 2020. Learning Unseen Concepts via Hierarchical Decomposition and Composition. In *CVPR*.

Yang, M.; Xu, C.; Wu, A.; and Deng, C. 2022. A Decomposable Causal View of Compositional Zero-Shot Learning. *IEEE TMM*, 1–11.

Yu, A.; and Grauman, K. 2014. Fine-Grained Visual Comparisons with Local Learning. In *CVPR*.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning Deep Features for Discriminative Localization. In *CVPR*.