A Decomposable Causal View of Compositional Zero-Shot Learning

Muli Yang, Chenghao Xu, Aming Wu, and Cheng Deng, *Senior Member, IEEE*

***Abstract*—Composing and recognizing novel concepts that are combinations of known concepts, *i.e.*, compositional generalization, is one of the greatest power of human intelligence. With the development of artificial intelligence, it becomes increasingly appealing to build a vision system that can generalize to unknown compositions based on restricted known knowledge, which has so far remained a great challenge to our community. In fact, machines can be easily misled by superficial correlations in the data, disregarding the causal patterns that are crucial to generalization. In this paper, we rethink compositional generalization with a causal perspective, upon the context of Compositional Zero-Shot Learning (CZSL). We develop a simple yet strong approach based on our novel Decomposable Causal view (dubbed "DECA"), by approximating the causal effect with the combination of three easy-to-learn components. Our proposed DECA[1] is evaluated on two challenging CZSL benchmarks by recognizing unknown compositions of known concepts. Despite being simple in the design, our approach achieves consistent improvements over state-of-the-art baselines, demonstrating its superiority towards the goal of compositional generalization.**

***Index Terms*—Compositional Zero-Shot Learning, Vision and Language, Image Recognition, Causality.**

I. INTRODUCTION

HUMANS are skilled at reasoning unknown concepts based on known knowledge. Imagining a *blue banana*, although most people have never seen one, they may immediately recognize a real blue banana when setting foot on Indonesia's Java Island. This ability, known as *compositional generalization* [1], is considered one ultimate goal of artificial intelligence, which is of great significance due to the compositional nature of our cognition system, *i.e.*, we build concepts on the combination of *primitives* [2]. In addition, compositional generalization is especially favored with limited supervision, given the fact that the long-tailed property [3] of real-world concepts makes it impractical to gather all possible combinations with full annotations.

In this work, we cast compositional generalization into the frame of Compositional Zero-Shot Learning (CZSL) [4], aiming to recognize novel combinations of known primitives. In CZSL, concepts are built on the combination of two primitives, namely, an attribute word and an object word, wherein training and inference happen in disjoint sets of these combinations. To infer unknown concepts such as *blue banana* in Fig. 1, a learning system needs to know how "blue" and a "banana" look like, after trained on other compositional concepts that separately contain "blue" and "banana", *e.g.*, *blue flower* and *ripe banana*. The challenge mainly lies in the entanglement of attributes and objects, which gives rise to varying *contextuality* within different attribute-object combinations [5]. As a result, semantic meanings of primitives are highly dependent on each other, leading to huge visual diversity that hinders the recognition of novel concepts.

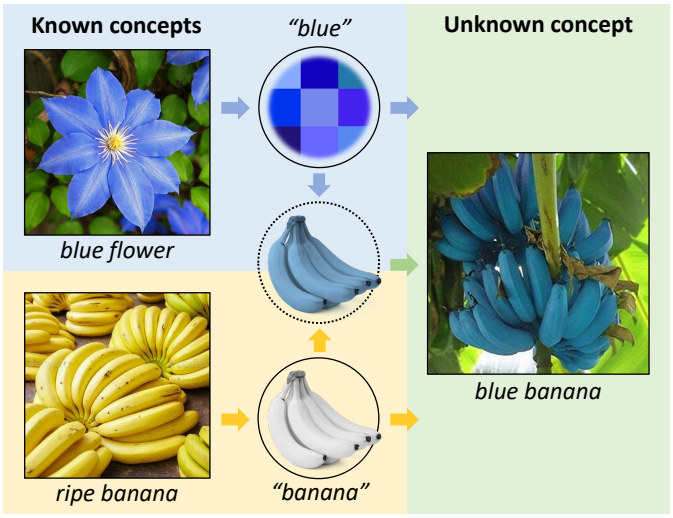


Fig. 1. An example of Compositional Zero-Shot Learning (CZSL) and the intuition of our proposed DECA. By training with some known concepts, such as *blue flower* and *ripe banana*, a model is expected to learn the primitives of the attribute, *i.e.*, *blue*, and the object, *i.e.*, *banana*, which can be used to compose and recognize the unknown concept — *blue banana*. In DECA, this is achieved by leveraging all three causal effects including attribute, object and composition effects, while most existing works only consider part of them.

Existing attempts mainly focus on separately modularizing attributes and objects [4], [5], or learning a shared embedding space for attribute-object compositions [6], [7]. In fact, independently treating attributes and objects disregards the contextuality between them, while solely modeling compositions as a whole actually imposes training-specific correlations that are detrimental to generalization. Despite being fairly effective, these methods suffer from the intractable contextuality, resulting in suboptimal compositional generalization.

In this paper, we rethink compositional generalization by asking: what do we actually need to recognize unknown

Manuscript received 16 February 2022; revised 20 June 2022; accepted 12 August 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62132016, Grant 62171343, and Grant 62071361, in part by the Key Research and Development Program of Shaanxi under Grant 2021ZDLGY01-03, and in part by the Fundamental Research Funds for the Central Universities ZDRC2102. *(Corresponding author: Cheng Deng.)*

Muli Yang, Chenghao Xu, Aming Wu, and Cheng Deng are with the School of Electronic Engineering, Xidian University, Xi'an 710071, China (e-mail: {mlyang, chx}@stu.xidian.edu.cn, amwu@xidian.edu.cn, chdeng@mail.xidian.edu.cn).

[1]Code is available on https://github.com/muliyangm/DeCa.

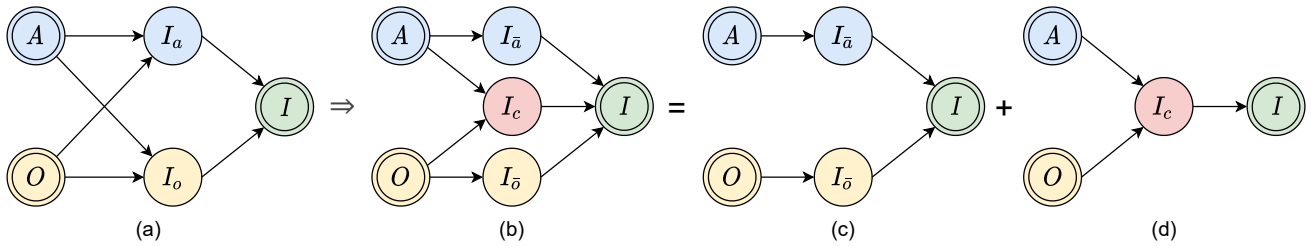


Fig. 2. **(a)** Our proposed causal graph for CZSL. Attribute A and object O simultaneously determine what image I may look like. The causal effect is passed through I_a and I_o, which respectively characterize the attribute and object effects to image I. Due to the *contextuality* between attributes and objects, I_a is not solely determined by A, but also by O; the same with I_o. It is thus not safe to directly disentangle I_a and I_o. Instead, as shown in **(b)**, we introduce I_c to capture the composition effect of both A and O, which allows pure attribute and object effects to be individually preserved in $I_{\bar{a}}$ and $I_{\bar{o}}$. Accordingly, in **(c)** and **(d)**, we learn the composition effect I_c and the attribute/object effects $I_{\bar{a}}, I_{\bar{o}}$ in a decomposable way, and then sum them for final predictions.

compositional concepts? We start with a causal view of how these concepts are formed, as shown in Fig. 2a. Intuitively, attributes and objects are two basic causes, whereas the contextuality renders them dependent, as discussed above. This dependence is characterized in Fig. 2a with two mediators I_a and I_o that have attributes A and objects O as common causes, reflecting their entangled nature. Therefore, coping with attributes and objects independently can be problematic as it violates this entangled nature. To remedy this, we propose to approximate the two dependent causal effects with three easy-to-compute components by introducing another mediator I_c to capture the entangled effect, shown in Fig. 2b, which facilitates learning intricate causal effects through a decomposable regime. Finally, the answer to our initial question is simple — all we need to know are how primitive concepts respectively look like, and once combined, how it may look like. This answer underlies the fact that a compositional concept means more than an attribute plus an object, but with much richer implications aroused by the contextuality. In fact, our decomposable causal view (DECA) actually serves as a more generalized CZSL solution encompassing most existing works that only consider part of the causal effects shown in Figs. 2c and 2d (see Sec. III-B).

To sum up, our main contributions are threefold:

- We propose a novel decomposable causal view of Compositional Zero-Shot Learning (CZSL), which underlies the problem and offers new insights to address CZSL;
- We develop an easy-to-implement pipeline tailored for learning decomposable causal effects in CZSL, which is compact in the structure and proven to be highly efficient;
- Extensive experiments and ablation studies verify the efficacy of our proposed method, which further validates the superiority of our decomposable causal view.

II. RELATED WORK

A. Compositional Generalization

Compositionality of visual attributes have been widely studied in recent years. These studies benefit the understanding of mid-level semantics that can be used to generalize to new compositions, *i.e.*, compositional generalization. Recent works have considered several tasks to evaluate the ability of compositional generalization, such as object detection [8], [9], action recognition [10], [11], and visual question answering [12]–[14]. In this paper, we focus on a special case of compositional generalization — Compositional Zero-Shot Learning (CZSL) [4] (detailed in Sec. II-C), and address it with a decomposable causal view.

B. Zero-Shot Learning (ZSL)

The aim of ZSL [15]–[17] 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 [18]–[24] aim to find a discriminative common embedding space for both visual features and attribute semantic features. Generating-based methods [25]–[29] 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.

C. Compositional Zero-Shot Learning (CZSL)

The aim of *Compositional Zero-Shot Learning (CZSL)* [4] is to infer unseen compositional concepts after training on some seen ones. As a specialized *Zero-Shot Learning (ZSL)* [17] task, CZSL shares the key ingredient with ZSL — exploiting transferable knowledge between seen and unseen concepts, which in ZSL is characterized by "attributes", *e.g.*, "stripes" for *tiger* and *zebra*. While in CZSL, such an "attribute" becomes a compositional part of the image label, *e.g.*, *wet dog*, where "wet" and "dog" can also be composed into novel concepts such as *wet cat* and *cute dog*. These compositions are often termed as "attribute-object" pairs for distinguishing the two compositional parts to be an "attribute" and an "object" [6]. In CZSL, only a portion of available compositions are involved in training, whereas all attributes and objects are seen during training, serving as a bridge to generalize to unseen compositions in testing.

Different from conventional ZSL [30], the key challenge of CZSL lies in the *contextuality* of attributes and objects [5], [31] — two compositional parts are highly-dependent to each

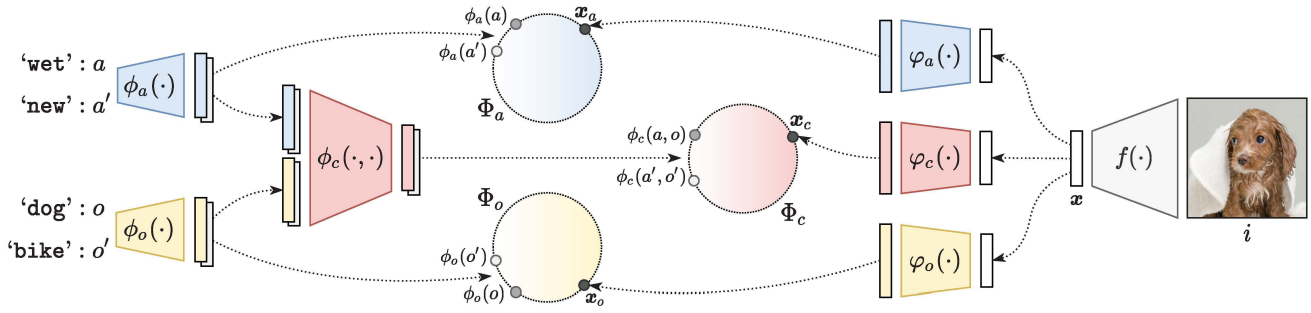


Fig. 3. Schematic of the proposed DECA. Taken the image of *wet dog* as an example, we embed its attribute *wet* and object *dog* into Φ_a and Φ_o as its attribute/object prototypes. All other attributes and objects in $\mathcal{A}$ and $\mathcal{O}$ are also embedded. Note that we only show each one of these (*e.g.*, *new* and *bike*) in this figure for brevity. Similarly, the composition *wet dog*, as well as other available compositions in $\mathcal{C}^s$, is embedded into Φ_c as the composition prototypes. As to the image, we first extract its feature and then map it into Φ_a, Φ_o, and Φ_c using three different transformations. In each embedding space, we regularize the image embedding to be close to its corresponding prototype while far from the others.

other. Due to a large number of possible compositions, such contextuality may scale into a huge generalization burden for novel compositions. In light of this, early attempts [4], [32], [33] often learn a transformation upon independent attribute and object classifiers to infer novel compositions. Another line of work proposes to model attributes as linear operators that modify the state of an object [6]; these operators are further regularized with a group of symmetric rules [7]. Similarly, manifold-based methods [31], [34], [35] directly map images and compositions into a shared embedding space with metric learning regularization, without consideration of the difference between attributes and objects. Some recent attempts [5], [36]–[41] focus on the compatibility between attributes, objects, and images that can be used to generalized to novel compositions. A most recent line of works [42], [43] employ self/cross-attention mechanism to model the relationships between attributes and objects. While in [44], the authors proposed to model CZSL with a causal graph, and impose independence regularization to learn disentangled representations of attributes and objects. A follow-up work [45] proposes a prototype propagation graph to learn compositional prototypes of novel attribute-object combinations. In this paper, similar to [44], we also resort to a causal view of CZSL, yet we provide a more generalized perspective that is totally different from the one in [44], which will be detailed in Sec. III-B.

D. Causality and Causal Inference

Causality describes the generic relationship between an effect and the cause that gives rise to it [46]–[48]. The study of causality between variables, namely causal inference, has been widely used in various real-world applications [49]–[51]. In recent years, researchers in machine learning communities also resort to causality in data [52] to learn more robust and interpretable models, *e.g.*, image recognition [53], [54], visual question answering [55], visual grounding [56], long-tailed classification [57], few-shot learning [58], zero-shot learning [59], and self-supervised learning [60].

In CZSL, a recent work [44] first proposes to address the problem within a causal framework, by assuming that attributes and objects are dependent to each other, leading to poor generalization to unknown compositions in inference. Accordingly, the key ingredient of this method is to disentangle the latent representations between attributes and objects. Considering the fact that contextuality plays an important role between attributes and objects within each composition, we argue that such disentanglement may be over-strict, leading to the loss of intra-composition knowledge that is beneficial to generalization. In this paper, we reconsider the causal relationships among attributes, objects, and compositions, and propose to model them using a decomposable causal graph with relaxed regularization compared to [44].

III. APPROACH

We present in this section our proposed approach to address CZSL. We first formally introduce the problem definition of CZSL, followed by our novel decomposable causal view (DECA) and its implementation. In the following, we use capital letters (*e.g.*, A) to denote random variables, and the lowercase ones as their observed values (*e.g.*, a).

A. Problem Definition

In CZSL, each image is composed of two primitive concepts, *i.e.*, an attribute (*e.g.*, *wet*) and an object (*e.g.*, *dog*). Given $\mathcal{A}$ and $\mathcal{O}$ as two sets of attributes and objects, we can compose a set of possible attribute-object pairs, *i.e.*, $\mathcal{C}=\mathcal{A}\times\mathcal{O}=\{(a,o)\,|\,a\in\mathcal{A},o\in\mathcal{O}\}$. Accordingly, we denote the training set as $\mathcal{D}^s=\{(i,c)\mid i\in\mathcal{I}^s,c\in\mathcal{C}^s\}$, in which $\mathcal{I}^s$ is an image set seen in training, and $\mathcal{C}^s$ is a subset of $\mathcal{C}$ containing the corresponding labels. In conventional Zero-Shot Learning, training and testing labels are non-overlapped, *i.e.*, $\mathcal{C}^s\cap\mathcal{C}^u=\varnothing$, where $\mathcal{C}^s,\mathcal{C}^u$ are two subsets of $\mathcal{C}$ seen/unseen in training. In this case, the model only needs to predict the compositions drawn from $\mathcal{C}^u$ in testing [4]. While in this paper, we follow the setting of generalized ZSL [61] where a testing sample can be drawn from either seen or unseen compositions ($\mathcal{C}^s\cup\mathcal{C}^u$), which is more challenging due to the larger prediction space and the dominant bias to seen compositions [5]. In a nutshell, the aim of CZSL is to learn a mapping $\mathcal{I}\mapsto\mathcal{C}^s\cup\mathcal{C}^u$ by training on $\{(\mathcal{I}^s,\mathcal{C}^s)\}$, in which $\mathcal{C}$ is composed of two primitive concepts drawn from $\mathcal{A}$ and $\mathcal{O}$.

B. A Decomposable Causal View (DECA)

In CZSL, models are required to estimate the likelihood $p(I=i \mid A=a, O=o)$ for image i conditioned on attribute a and object o; for inference, we calculate

$$\hat{c} = (\hat{a}, \hat{o}) = \underset{(a,o)\in\mathcal{C}}{\arg\max}\ p(i \mid a, o) \tag{1}$$

as the prediction result for image i.

To describe the cause-effect relationship in CZSL, we follow [44] to treat labels as causes of images, rather than their effects, which underlies the image generation process and also enables unseen composition recognition. We present in Fig. 2a our proposed causal graph, where we use mediators I_a and I_o to pass the effects of A and O to I, similar to [44]. This is achieved by learning two mappings from $\mathcal{A}$ and $\mathcal{O}$ to attribute/object embedding spaces Φ_a and Φ_o, where each attribute/object prototype is estimated from the images, measured by $p(i_a \mid a)$ and $p(i_o \mid o)$, for i_a and i_o are two representations encoding the attribute/object seen in image i.

To address *contextuality* between A and O, in [44] there is a two-way causal relationship between attribute A and object O, while in Fig. 2a we deem that A and O are not necessarily correlated at the beginning, but instead entangled in mediators I_a and I_o as they co-determine what image I may look like. Taken *old dog* and *old car* as examples, their common attribute *old* can hardly be characterized without the object as it appears distinctively when associated with different objects — an *old dog* tends to have thinner and duller hair with cloudy eyes, while an *old car* may be an old-fashioned one covered with rust. This phenomenon also holds for objects, such as *ripe apple* and *sliced apple*, in which a same object presents distinct appearance under different attributes. In these cases, it is extremely difficult to disentangle attributes from objects, and vice versa, due to the intrinsic contextuality between them. In light of this, each mediator (I_a or I_o) is designed to respond to both the attribute A and the object O. This design respects the contextuality between A and O, and is free from the explicit disentanglement between them, characterized as paths $A, O \rightarrow I_a \rightarrow I$ and $A, O \rightarrow I_o \rightarrow I$ in Fig. 2a.

It is tricky to model the causal effects using machine learning building blocks given the current form of causal graph in Fig. 2a. A possible option is to disentangle I_a from I_o with independence regularization on Φ_a and Φ_o [44]. However, a direct disentanglement actually violates the correlative nature of I_a and I_o since they both have A and O as common causes. In view of this, we divide the mediator I_a into two parts, *i.e.*, $I_{\bar{a}}$ and I_{ac}, which respectively record the pure effect of attribute A (*e.g.*, the faded atmosphere of the attribute *old*) and the effect of A influenced by the contextuality (*e.g.*, the dull hair of an *old dog* and the rusty surface of an *old car*). Likewise, I_o is also divided into $I_{\bar{o}}$ and I_{oc}. Ideally, the two pure effects $I_{\bar{a}}$ and $I_{\bar{o}}$ should be independent since they are free from the influence of contextuality. Accordingly, as shown in Fig. 2b, we introduce another mediator I_c to address the contextuality by merging I_{ac} and I_{oc} into a whole since each of them is determined by both A and O. The mediator I_c captures the composition effect of A and O, *i.e.*, $A, O \rightarrow I_c \rightarrow I$, allowing the individual pure effects of A and O to be better preserved in $I_{\bar{a}}$ and $I_{\bar{o}}$. To achieve this, we learn another mapping $\mathcal{C} \mapsto \Phi_c$ in addition to the existing $\mathcal{A} \mapsto \Phi_a$ and $\mathcal{O} \mapsto \Phi_o$, and rewrite the corresponding likelihood as $p(i_c \mid a, o)$, $p(i_{\bar{a}} \mid a)$, and $p(i_{\bar{o}} \mid o)$, in which i_c, $i_{\bar{a}}$, and $i_{\bar{o}}$ are three different representations of image i.

Algorithm 1: Training procedure of DECA.

Data: Training data $\mathcal{D}^s$, temperature parameter τ
Result: Optimal $\varphi_a(\cdot)$, $\varphi_o(\cdot)$, $\varphi_c(\cdot)$, $\phi_a(\cdot)$, $\phi_o(\cdot)$, $\phi_c(\cdot,\cdot)$
Initialize: $f(\cdot)$, $\varphi_a(\cdot)$, $\varphi_o(\cdot)$, $\varphi_c(\cdot)$, $\phi_a(\cdot)$, $\phi_o(\cdot)$, $\phi_c(\cdot,\cdot)$
while *not converged* **do**
  Sample a batch from $\mathcal{D}^s$ as images $\{i_k\}_{k=1}^n$ with their labels (attributes/objects) $\{(a_k, o_k)\}_{k=1}^n$;
  for *samples in the batch* **do**
    Extract visual features: $\boldsymbol{x} = f(i)$;
    Map visual features to three embedding spaces: $\boldsymbol{x}_a = \varphi_a(\boldsymbol{x})$, $\boldsymbol{x}_o = \varphi_o(\boldsymbol{x})$, $\boldsymbol{x}_c = \varphi_c(\boldsymbol{x})$;
    Map attributes/objects to three embedding spaces: $\boldsymbol{v}_a = \phi_a(a)$, $\boldsymbol{v}_o = \phi_o(o)$, $\boldsymbol{v}_c = \phi_c(a, o)$;
    Calculate $\mathcal{L}_a$, $\mathcal{L}_o$, and $\mathcal{L}_c$ based on Eqs. (3) to (5);
  end
  $\mathcal{L} = 0.5 \times (\mathcal{L}_a + \mathcal{L}_o) + \mathcal{L}_c$;
  Update network parameters using $\nabla\mathcal{L}$;
end

The transformed causal graph in Fig. 2b observes the *contextuality* between A and O, which also serves as an easy-to-implement proxy for the original graph in Fig. 2a. Benefiting from its decomposable structure, we can easily estimate the causal effects in three different embedding spaces (Φ_a, Φ_o, and Φ_c) at two hierarchies as shown in Figs. 2c and 2d. Consequently, the final prediction is derived by summing the three effects, and we can rewrite the inference rule (which also holds in training) in Eq. (1) as

$$(\hat{a}, \hat{o}) = \underset{(a,o)\in\mathcal{C}}{\arg\max}\ p(i_{\bar{a}} \mid a)\, p(i_{\bar{o}} \mid o)\, p(i_c \mid a, o)\,. \tag{2}$$

With our decomposable causal view, most existing works can actually be regarded as a special case of our proposed solution, which only consider part of the three causal effects, *i.e.*, either only attribute + object effects (Fig. 2c): [7], [44], or only composition effect (Fig. 2d): [4]–[6], [35], [41]. We will show in Sec. IV-B that solely considering part of the causal effects leads to suboptimal results — only by obeying our decomposable causal view to use all the three causal effects can we achieve the optimal.

C. Implementation

1) Parameterization: Our whole pipeline is illustrated in Fig. 3. The three mappings, *i.e.*, $\mathcal{A} \mapsto \Phi_a$, $\mathcal{O} \mapsto \Phi_o$, and $\mathcal{C} \mapsto \Phi_c$, are parameterized by three neural networks $\phi_a(\cdot)$, $\phi_o(\cdot)$, and $\phi_c(\cdot,\cdot)$. Accordingly, we have $\boldsymbol{v}_a = \phi_a(a) \in \Phi_a$, $\boldsymbol{v}_o = \phi_o(o) \in \Phi_o$, and $\boldsymbol{v}_c = \phi_c(a, o) \in \Phi_c$ as embedding vectors in each embedding space. Notably, these embedding vectors serve as prototypes that semantically characterize attribute/object/composition representations.

For image i, we first extract its visual feature using a standard convolutional neural network $f(\cdot)$, which maps $\mathcal{I}$ into the visual feature space $\mathcal{X}$, *i.e.*, $\boldsymbol{x} = f(i)$. Naturally, we expect that the observed attribute in image i would be

TABLE I
DATASET DETAILS WITH RESPECT TO ATTRIBUTE/OBJECT NUMBERS, PAIR/IMAGE NUMBERS IN SEEN/UNSEEN SPLITS, AND IN VAL/TEST SETS.

Subset →			Seen (S)		Unseen (U)		Val Set		Test Set	
Dataset ↓	Attr.	Obj.	Pairs	Images	Pairs	Images	Pairs (S/U)	Images (S/U)	Pairs (S/U)	Images (S/U)
MIT-States [62]	115	245	1,262	34,562	700	19,191	300/300	1,844/8,576	400/400	2,380/10,615
UT-Zappos [63]	16	12	83	24,898	33	4,228	15/15	877/2,337	18/18	1,023/1,891

embedded close to its attribute prototype $\phi_a(a)$; the same with its object and composition prototypes $\phi_o(o)$, $\phi_c(a,o)$. To do so, we need three additional mappings $\mathcal{X} \mapsto \Phi_a, \Phi_o, \Phi_c$ that further map the visual feature $\boldsymbol{x}$ into attribute, object, and composition embedding spaces, *i.e.*, $\boldsymbol{x}_a = \varphi_a(\boldsymbol{x})$, $\boldsymbol{x}_o = \varphi_o(\boldsymbol{x})$, and $\boldsymbol{x}_c = \varphi_c(\boldsymbol{x})$.

2) Training: Our training objective is to maximize the three likelihood discussed in Sec. III-B, *i.e.*, $p(i_{\bar{a}} \,|\, a)$, $p(i_{\bar{o}} \,|\, o)$, and $p(i_c \mid a, o)$. With the above parameterization, this objective can be achieved by ensuring that each image is embedded closest to its three prototypes in the embedding spaces Φ_a, Φ_o, and Φ_c. To this end, we adopt an InfoNCE-like loss function that is widely-used in contrastive learning literature [64]–[66]. However, we highlight that our formulation is different from conventional contrastive learning which relies on strong augmentations to achieve transformation consistency [67]–[69]. On the contrary, ours is augmentation-free by treating all attribute/object/composition prototypes as positive/negative samples to each image in the three embedding spaces, which enables fast convergence within an easy-to-implement formulation. It is worth noting that our loss is in form similar to that in CompCos [35], yet is motivated from a different perspective. Our loss is designed to regularize the correlations between visual features and the semantic prototypes in the three embedding spaces at two levels: 1) the individual level and 2) the composition level (Fig. 3). This design not only respects the independence between the pure effects of attributes and objects by separately constructing their prototypes, but also addresses the contextuality by constructing the composition prototype based on both attribute and object prototypes.

Specifically, given an image-label pair (i, c), or $(i, (a, o))$, in training set $\mathcal{D}^s$, we first embed the image into the three embedding spaces as $\boldsymbol{x}_a, \boldsymbol{x}_o, \boldsymbol{x}_c$. At the same time, all attributes, objects, and compositions in $\mathcal{A}$, $\mathcal{O}$, and $\mathcal{C}^s$ are embedded into the three spaces as attribute/object/composition prototypes. For attribute and object embeddings, we minimize the following losses to ensure the largest similarities between $\boldsymbol{x}_a, \boldsymbol{x}_o$ and their corresponding attribute/object prototypes $\phi_a(a), \phi_o(o)$:

$$\mathcal{L}_a = -\log \frac{\exp\big(\langle \boldsymbol{x}_a, \phi_a(a)\rangle/\tau\big)}{\sum_{a' \in \mathcal{A}} \exp\big(\langle \boldsymbol{x}_a, \phi_a(a')\rangle/\tau\big)}\,, \tag{3}$$

$$\mathcal{L}_o = -\log \frac{\exp\big(\langle \boldsymbol{x}_o, \phi_o(o)\rangle/\tau\big)}{\sum_{o' \in \mathcal{O}} \exp\big(\langle \boldsymbol{x}_o, \phi_o(o')\rangle/\tau\big)}\,, \tag{4}$$

where τ is a temperature parameter [70] that balances the losses by scaling the model output, and $\langle \cdot, \cdot \rangle$ represents the similarity of two vectors; in this work we respectively experiment on dot product, *i.e.*, $\langle \boldsymbol{x}, \boldsymbol{a} \rangle = \boldsymbol{x} \cdot \boldsymbol{a} = \boldsymbol{x}\boldsymbol{a}^\mathsf{T}$, and cosine similarity, *i.e.*, $\langle \boldsymbol{x}, \boldsymbol{a} \rangle = \boldsymbol{x} \cdot \boldsymbol{a} / \|\boldsymbol{x}\| \|\boldsymbol{a}\|$. Similarly, we have

$$\mathcal{L}_c = -\log \frac{\exp\big(\langle \boldsymbol{x}_c, \phi_c(a,o)\rangle/\tau\big)}{\sum_{(a',o') \in \mathcal{C}^s} \exp\big(\langle \boldsymbol{x}_c, \phi_c(a',o')\rangle/\tau\big)}\,, \tag{5}$$

which ensures $\boldsymbol{x}_c$ to be closest to its corresponding composition prototype $\phi_c(a, o)$. Finally, we linearly combine the above losses as the training objective for DECA:

$$\mathcal{L} = 0.5 \times (\mathcal{L}_a + \mathcal{L}_o) + \mathcal{L}_c\,. \tag{6}$$

As discussed earlier in Sec. III-B, $I_{\bar{a}}$ and $I_{\bar{o}}$ in Fig. 2c should be independent as they capture the pure individual effects free from the influence of contextuality. Our proposed DECA does not involve any explicit regularization to guarantee this independence as in prior works Causal [44] and Proto-Prop [45]. In fact, we have also tried such extra regularization, yet with no obvious benefit (also no obvious harm). We hypothesize that our separate modularization of the three causal effects already implicitly promotes the independence, due to the intrinsic difference between the learned manifolds of the embedding spaces under distinct training objectives, which encourage an implicit disentanglement among attributes, objects, and compositions by aligning different visual features with different semantic prototypes in different embedding spaces.

3) Inference: The inference takes place in both $\mathcal{C}^s$ and $\mathcal{C}^u$. Similar to in training, we embed an image as $\boldsymbol{x}_a, \boldsymbol{x}_o, \boldsymbol{x}_c$, and compare them with all available attribute/object/composition prototypes mapped from $\mathcal{A}$, $\mathcal{O}$, and $\mathcal{C}$; then the attribute and object of the most similar prototypes are taken as the prediction. The inference rule, also depicted in Eq. (2), can be parameterized as

$$\begin{aligned}(\hat{a}, \hat{o}) = \underset{(a,o) \in \mathcal{C}}{\arg\max}\; & \alpha\big(\langle \boldsymbol{x}_a, \phi_a(a)\rangle + \langle \boldsymbol{x}_o, \phi_o(o)\rangle\big) \\ & + (1-\alpha)\langle \boldsymbol{x}_c, \phi_c(a,o)\rangle\,,\end{aligned} \tag{7}$$

where $\alpha \in [0, 1]$ trades off between the two parts of causal effects (Figs. 2c and 2d) in inference.

IV. EXPERIMENTS

A. Experimental Setup

1) Datasets: Our proposed DECA is evaluated on two CZSL benchmark datasets, *i.e.*, MIT-States [62] and UT-Zappos [63].

MIT-States contains 53,753 everyday images, *e.g.*, "cute cat" and "tall building", with 115 attributes and 245 objects in total. MIT-States has 1962 available attribute-object pairs, in which 1262 pairs are seen in training, leaving the other 700 pairs unseen. *UT-Zappos* contains 50,025 images of shoes, *e.g.*, "canvas slippers" and "rubber sandals", with 16 attributes and

TABLE II
ABLATION STUDY. RESULTS ARE REPORTED IN SEEN/UNSEEN COMPOSITION RECOGNITION ACCURACY (%) IN VALIDATION SETS OF MIT-STATES AND UT-ZAPPOS. **BEST** AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN EACH COLUMN.

Dataset →		MIT-States						UT-Zappos					
#	Method ↓	Attr.	Obj.	Seen	Unseen	HM	AUC	Attr.	Obj.	Seen	Unseen	HM	AUC
(1)	w/o Φ_a	30.1	**36.5**	30.1	28.5	20.7	6.6	49.4	**73.4**	60.0	65.7	55.1	37.1
(2)	w/o Φ_o	31.8	33.5	30.0	27.7	19.9	6.3	**62.4**	63.4	61.6	63.7	45.9	31.1
(3)	w/o Φ_c	29.7	32.5	26.0	25.2	17.7	4.9	52.8	73.2	**62.1**	67.4	51.9	36.4
(4)	w/o Prior	31.8	36.1	**31.2**	28.6	20.9	6.9	55.0	73.1	61.8	**70.2**	55.8	39.6
(5)	Dot Product	30.8	35.9	30.3	28.6	20.2	6.6	61.7	71.3	61.0	69.5	53.5	37.9
(6)	Cosine (Full DECA)	**32.1**	36.1	30.6	**28.9**	**21.0**	**6.9**	56.2	71.8	61.5	69.5	**55.8**	**39.7**

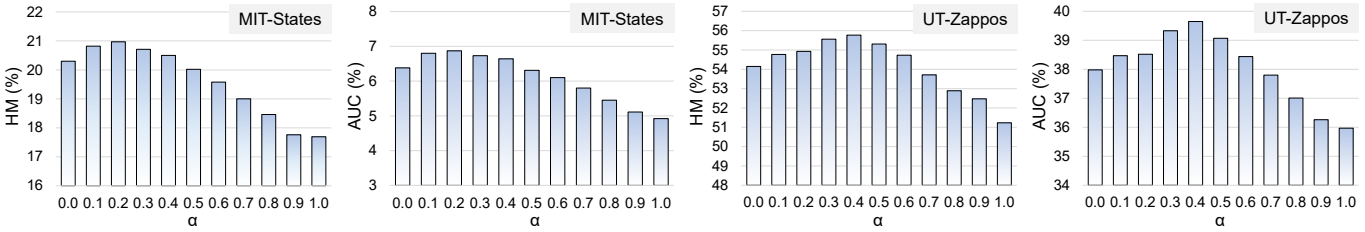


Fig. 4. The impact of α in Eq. (7). Results are reported in seen/unseen composition recognition accuracy (%).

12 objects. In UT-Zappos there are 116 attribute-object pairs, 83 pairs of which are used for training, while the other 33 pairs are unseen in training.

We follow the widely-adopted protocol [5] to prepare train/-val/test splits. For MIT-States, we use 30,338 images from 1262 seen pairs for training, 10,420 images from 300 seen and 300 unseen pairs for validation, and 12,995 images from 400 seen and 400 unseen pairs for test. For UT-Zappos, the training set contains 22,998 images from 83 seen pairs; the validation set contains 3214 images from 15 seen and 15 unseen pairs; the test set contains 2914 images from 18 seen and 18 unseen pairs. The dataset details are summarized in Tab. I.

2) Evaluation Metrics: We evaluate the performance according to prediction accuracy for recognizing seen and unseen attribute-object pairs. Following [5], we compute the best accuracy in two situations, namely, 1) *Seen*, only testing on seen pairs; 2) *Unseen*, only testing on unseen pairs. Accordingly, we can compute 3) *Harmonic Mean (HM)* of the two metrics:

$$Acc_{\text{HM}} = 2 \times \frac{Acc_{\text{Seen}} \times Acc_{\text{Unseen}}}{Acc_{\text{Seen}} + Acc_{\text{Unseen}}}, \tag{8}$$

which balances the performance between seen and unseen accuracy. Finally, we compute 4) *Area Under the Curve (AUC)* to quantify the overall performance of both seen and unseen accuracy. Following [5], [61], we use a calibration bias to trade off between the prediction scores of seen and unseen pairs; as the calibration bias varies, we can draw a seen-unseen accuracy curve on which the AUC metric can be computed.

3) Implementation Details: We summarize our core implementation in Algorithm 1. For the image encoder $f(\cdot)$, we follow [5], [7], [35] to use a ResNet-18 [71] pre-trained on ImageNet [72] without fine-tuning for fair comparison with the prior work. We also provide in Tabs. III and IV the results with the image encoder fine-tuned for comprehensive evaluations over state-of-the-art competitors. To achieve that, we reimplement the involved methods using the official code with the optimal hyperparameters, and report the results averaged over three random seeds.

The three visual mappings $\varphi_a(\cdot)$, $\varphi_o(\cdot)$, and $\varphi_c(\cdot)$ are implemented as two-layer fully-connected networks on top of the image encoder $f(\cdot)$, with LayerNorm [73], ReLU [74] activation and Dropout [75] applied after the first layer. Attribute/object mappings $\phi_a(\cdot)$ and $\phi_o(\cdot)$ are implemented as word embedding functions. Following [35], [41], the word embeddings are initialized with 300-dimensional Word2Vec [76] embeddings for UT-Zappos, and 600-dimensional Word2Vec+fastText [77] embeddings for MIT-States. For mapping compositions, *i.e.*, $\phi_c(\cdot)$, we first obtain attribute and object embeddings (*i.e.*, $\phi_a(a)$ and $\phi_o(o)$) for each composition, and then feed their concatenation to a fully-connected layer that reduces the dimension to its half, *i.e.*, the same with other embeddings. Note that a (rationally) larger embedding dimension almost always means better recognition performance, and thus we strictly follow prior work [35], [41] to keep the dimension intact for both visual and linguistic embeddings, *i.e.*, 600 for MIT-States and 300 for UT-Zappos.

The whole model, except the fixed image encoder, is trained using Adam optimizer [78] with batch size of 256, learning rate of $5{\times}10^{-4}$, and weight decay of $5{\times}10^{-5}$, within 50 epochs. When using cosine similarity for $\langle \cdot, \cdot \rangle$ in Eqs. (3) to (5), the temperature parameter τ is set to 0.05 and 0.02 for MIT-States and UT-Zappos, respectively; when using dot product, we simply set $\tau = 1$ for both datasets. Unless otherwise mentioned, we use cosine similarity by default. In inference, α in Eq. (7) is respectively set to 0.2 and 0.4 for MIT-States and UT-Zappos. Our model is implemented using PyTorch on an NVIDIA TITAN Xp GPU, and the code is available on https://github.com/muliyangm/DeCa.

TABLE III
COMPARISON WITH STATE-OF-THE-ART BASELINES IN **MIT-STATES**. RESULTS ARE REPORTED IN SEEN/UNSEEN COMPOSITION RECOGNITION ACCURACY (%). †FINE-TUNING THE IMAGE ENCODER DURING TRAINING. **BEST** AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN EACH COLUMN.

Protocol →		Val AUC			Test AUC			Test		
Method ↓	Venue ↓	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3	Seen	Unseen	HM
RedWine [4]	CVPR'17	2.9	7.3	11.8	2.4	5.7	9.3	20.7	17.9	11.6
AttrAsOp [6]	ECCV'18	2.5	6.2	10.1	1.6	4.7	7.6	14.3	17.4	9.9
LabelEmbed+ [6]	ECCV'18	3.0	7.6	12.2	2.0	5.6	9.4	15.0	20.1	10.7
TMN [5]	ICCV'19	3.5	8.1	12.4	2.9	7.1	11.5	20.2	20.1	13.0
SymNet [7]	CVPR'20	4.3	9.8	14.8	3.0	7.6	12.3	24.4	25.2	16.1
Causal [44]	NeurIPS'20	1.7	4.0	5.9	1.5	3.4	5.3	17.5	11.8	9.5
CGE [41]	CVPR'21	6.8	–	–	5.1	–	–	28.7	**25.3**	17.2
CompCos [35]	CVPR'21	5.9	13.4	19.8	4.5	10.9	16.5	25.3	24.6	16.4
ProtoProp [45]	NeurIPS'21	4.1	9.5	14.4	2.7	7.0	11.3	19.2	20.4	12.6
DECA (Ours)	–	**6.9**	**14.9**	**21.6**	**5.3**	**12.4**	**18.5**	**29.8**	25.2	**18.2**
Causal† [44]	NeurIPS'20	2.4	4.9	6.7	1.9	4.3	6.5	19.5	13.2	10.6
CGE† [41]	CVPR'21	8.2	**17.2**	24.2	6.5	14.3	20.6	**32.9**	27.1	20.0
CompCos† [35]	CVPR'21	7.3	15.6	21.9	5.7	12.8	18.7	29.4	26.8	18.8
ProtoProp† [45]	NeurIPS'21	4.7	10.4	15.2	3.1	7.8	11.9	23.9	18.8	13.7
DECA† (Ours)	–	**8.3**	17.1	**24.2**	**6.6**	**14.7**	**22.2**	32.2	**27.4**	**20.3**

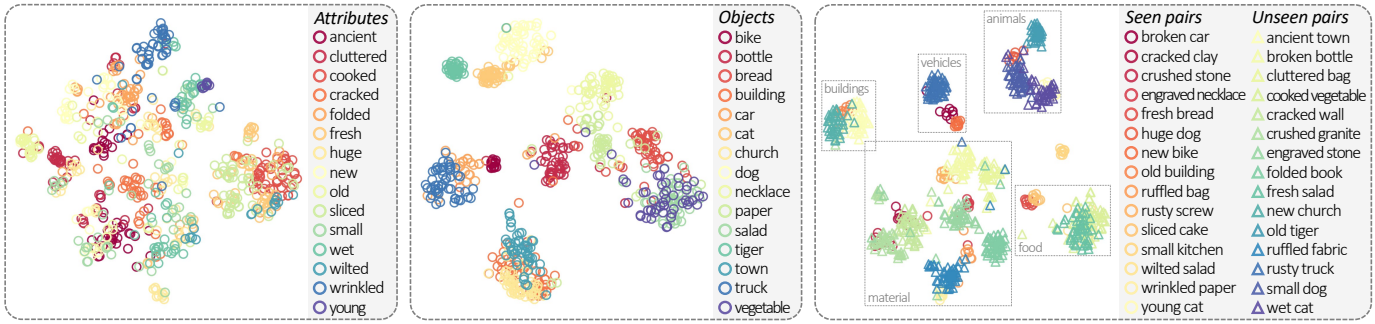
TABLE IV
COMPARISON WITH STATE-OF-THE-ART BASELINES IN **UT-ZAPPOS**. RESULTS ARE REPORTED IN SEEN/UNSEEN COMPOSITION RECOGNITION ACCURACY (%). †FINE-TUNING THE IMAGE ENCODER DURING TRAINING. **BEST** AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN EACH COLUMN.

Protocol →		Val AUC			Test AUC			Test		
Method ↓	Venue ↓	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3	Seen	Unseen	HM
RedWine [4]	CVPR'17	30.4	52.2	63.5	27.1	54.6	68.8	57.3	62.3	41.0
AttrAsOp [6]	ECCV'18	21.5	44.2	61.6	25.9	51.3	67.6	59.8	54.2	40.8
LabelEmbed+ [6]	ECCV'18	26.4	49.0	66.1	25.7	52.1	67.8	53.0	61.9	40.6
TMN [5]	ICCV'19	36.8	57.1	69.2	29.3	55.3	69.8	58.7	60.0	45.0
SymNet [7]	CVPR'20	25.9	–	–	23.4	–	–	49.8	57.4	40.4
Causal [44]	NeurIPS'20	21.0	43.4	58.3	24.3	47.1	62.0	59.1	51.8	40.5
CGE [41]	CVPR'21	38.7	–	–	26.4	–	–	56.8	**63.6**	41.2
CompCos [35]	CVPR'21	38.6	60.1	71.8	28.7	55.9	72.5	59.8	62.5	43.1
ProtoProp [45]	NeurIPS'21	31.6	52.0	65.5	23.2	47.3	63.2	54.1	54.7	38.8
DECA (Ours)	–	**39.7**	**61.5**	**73.6**	**31.6**	**58.2**	**73.7**	**62.7**	63.1	**46.3**
Causal† [44]	NeurIPS'20	30.8	54.2	63.9	26.7	51.2	65.3	62.2	53.4	43.5
CGE† [41]	CVPR'21	41.8	64.1	77.4	34.7	64.7	78.1	62.6	67.8	49.5
CompCos† [35]	CVPR'21	42.4	66.0	77.4	35.2	64.9	78.4	**64.5**	67.5	49.5
ProtoProp† [45]	NeurIPS'21	39.6	61.9	75.2	34.2	60.3	74.1	61.0	64.9	50.1
DECA† (Ours)	–	**42.8**	**66.1**	**78.2**	**37.0**	**65.2**	**78.6**	64.0	**68.8**	**51.7**

B. Ablation Study

We ablate our DECA to evaluate the effectiveness of each proposed module. The ablation study is conducted in the validation set of both datasets — we only use the test set for comparison with state of the arts when parameter tuning is done in the val set. The results are summarized in Tab. II and Fig. 4, *w.r.t.* the following two aspects.

1) Embedding Spaces: We first study the effects of Φ_a, Φ_o, and Φ_c, corresponding to (1)–(3) in Tab. II. Specifically, we exclude a certain embedding space in both training and inference, and report in Tab. II the results in terms of six evaluation metrics, with additional attribute/object classification accuracy apart from the four metrics introduced in Sec. IV-A2. Compared to our full model (6), all three embedding spaces contribute to compositional recognition results, indicating the necessity to incorporate all three causal effects in inference. As to the three embedding spaces, from (2) and (3) we can observe that Φ_o is more important in UT-Zappos, while the same case is Φ_c in MIT-States. This is mainly because of the distribution difference between the two datasets: The images in MIT-States present significant visual diversity, in terms of different backgrounds, postures, lighting, *etc*., while UT-Zappos contains only same-orientated shoe images with white background; further, MIT-States contains much more attributes (115) and objects (245) while the statistic of UT-Zappos is 16 and 12. As a result, it can be difficult to learn accurate object prototypes in MIT-States, while this is not the case for UT-Zappos. Due to the large visual diversity of attributes, Φ_a seems to be overall less advantageous, yet still, far from negligible. In general, our decomposable causal view takes



(a) Attribute space Φ_a (b) Object space Φ_o (c) Compositional space Φ_c

Fig. 5. t-SNE [79] visualizations of attribute/object/compositional embedding spaces. The visualized samples are all from the test set of MIT-States, marked with various colors for different attributes/objects/compositions, and different shapes for seen/unseen compositions. Best viewed in color.

all considerations of the factors needed to recognize compositional concepts, and thus is less sensitive to the distribution difference in the datasets, as we will further show in Sec. IV-C.

We also report in (4) the results without prior word embedding (*e.g.*, Word2Vec, discussed in Sec. IV-A3) to initialize attribute and object mappings, *i.e.*, $\phi_a(\cdot)$ and $\phi_o(\cdot)$. It shows that our method is able to learn promising embeddings without the prior guidance that is requisite for many former approaches [5], [41].

Additionally, we show in (5) and (6) that a cosine similarity is slightly better than dot product in Eqs. (3) to (5) for learning the embedding spaces. This is also observed in other low-shot learning literature [80], [81] that a normalized similarity generalizes better to unfamiliar distributions.

It is worth noting that our full model in (6) does not always achieve the best performance in single attribute/object concepts, especially in UT-Zappos. Due to the existence of spurious correlations, a naïve baseline can easily overfit to single concepts, achieving superior performance in classifying attributes (objects), yet manifesting poor performance in objects (attributes). In contrast, the proposed DECA is designed to tune down the causal effect of single concepts by incorporating all three effects of attributes, objects, and compositions. This design may decrease the classification accuracy of single concepts, yet is helpful when classifying both attributes and objects, resulting in superior performance in CZSL.

2) Inference Rule: We study the impact of α in Eq. (7) after our full model is trained. Concretely, α controls the proportion of causal effects to use in inference, *i.e.*, how much we should trust attribute, object, and composition effects, respectively. In this work, we equally treat attribute and object effects for brevity, despite that carefully tuning their proportion will arguably lead to better accuracy.

As two extreme cases, when $\alpha=0$, only composition effect is used in inference; when $\alpha=1$, we only use the sum of attribute and object effects. We report in Fig. 4 the results when $\alpha \in [0,1]$ with a 0.1 interval. As can be observed, in MIT-States, the recognition accuracy reaches its peak in terms of both HM and AUC metrics when α equals 0.2; in UT-Zappos, the value is 0.4. A possible explanation is that the composition effect provides more precise decision signal than the attribute/object effect since only a small fraction of all possible attribute-object combinations are involved in MIT-States as compositional concepts; in contrast, UT-Zappos contains most of the possible combinations, such that the side information provided by the attribute/object effect can be of more help to compositional generalization.

Another important observation is that solely using the composition effect ($\alpha=0$) is always better than using the sum of attribute and object effects (α=1). The main reason is that, the goal of CZSL is to correctly classify compositional concepts (corresponding to the composition effect) rather than single attribute/object concepts (corresponding to the attribute/object effect), while separately modeling them disregards their contextuality that is crucial to compositional generalization. Still, using the three effects altogether reaches the optimal, which incorporates the contextuality as well as the unique primitive knowledge that is beneficial to generalization.

C. Comparison with State of the Arts

We compare our proposed DECA with nine state-of-the-art baselines, which are introduced below.

1) RedWine [4] trains linear SVMs for attribute/object primitives and then transforms their weight parameters using a neural network for recognizing unseen compositions.

2) AttrAsOp [6] regards attributes as operators, modeling compositions as attribute-guided transformations of objects.

3) LabelEmbed+ [6] concatenates attributes and objects as compositions after embedding them using GloVe [82].

4) TMN [5] proposes modularized networks by dividing the task of recognizing unseen compositions into sub-tasks.

5) SymNet [7] resorts to the group theory by leveraging the symmetric principles when composing attribute-object pairs.

6) Causal [44] learns disentangled representations between attributes and objects inspired by a causal view.

7) CGE [41] uses Graph Neural Networks to model the long-term dependence between attributes and objects.

8) CompCos [35] directly maps the attribute-object compositions and visual features into a common embedding space.

9) ProtoProp [45] proposes a prototype propagation graph method with independent constraints on attributes and objects.

We report the results in Tabs. III and IV, respectively for MIT-States and UT-Zappos, from which we can observe that our proposed DECA consistently outperforms all baselines in

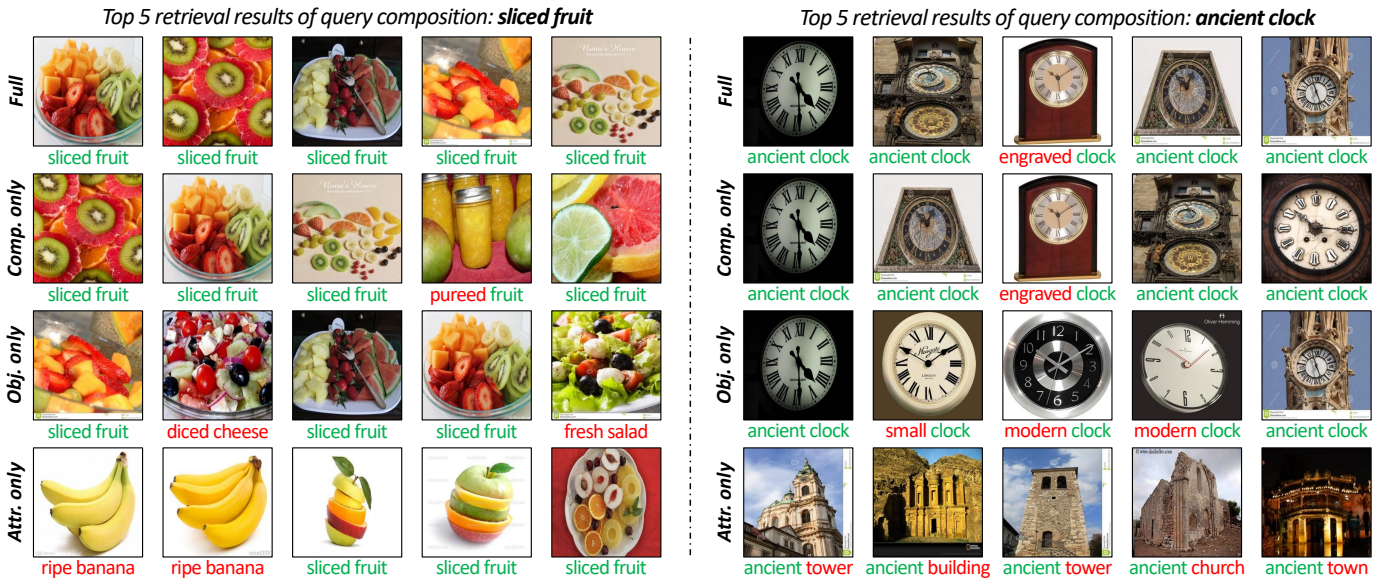


Fig. 6. Retrieval results of unseen compositions from test set images of MIT-States. We show retrieved images of the same compositions using different inference rules. Correct and incorrect primitives are marked in green and red under each retrieved image, respectively. Best viewed in color.

terms of both HM and AUC metrics. Notably, most baselines are sensitive to data distributions, *i.e.*, only performing favorably in either MIT-States or UT-Zappos. For instance, TMN [5] achieves the highest test AUC of 29.3 among all baselines in UT-Zappos, while only reaching 2.9 in MIT-States, less than two thirds of that of CGE [41], the best baseline in MIT-States. This also occurs in the newer baselines, *e.g.*, SymNet [7] and CGE, both favoring MIT-States. As to Causal [44] and ProtoProp [45], they both show competitive performance in UT-Zappos yet fail to produce favorable results in MIT-States. The reason can be the huge difference between the two datasets: MIT-States is a large-scale real-world dataset with thousands of compositional concepts, while UT-Zappos is a much simpler one with 100+ compositions of shoe images that are less ambiguous, as discussed earlier in Sec. IV-B. In contrast, the proposed DECA performs well in both datasets, demonstrating the superiority and broad applicability of our decomposable causal view.

D. Qualitative Evaluations

1) Embedding Space Visualizations: We show in Fig. 5 the t-SNE [79] visualizations of the three embedding spaces. Considering the huge amount of available concepts (*cf.* Tab. I), we randomly choose subsets of attributes/objects/compositions with 1) a broad concept span and 2) sufficient image samples for clear visualizations of the three embedding spaces. In Fig. 5a, due to the ambiguous nature of attributes, the embedding space seems to be cluttered with small clusters, while in Fig. 5b we can observe clearer clusters since objects, corresponding to physical entities, are more separable than attributes. Finally, in Fig. 5c, samples are tightly clustered since they are directly labeled as compositions, which most accurately characterize the data points other than single attributes or objects; we can also observe that similar concepts are grouped together, roughly forming super categories such as buildings and animals. Still, there exist hard-to-separate samples within the compositional space in Fig. 5c; by leveraging all three causal effects we can better recognize unseen compositions (*cf.* Tab. II) benefiting from the complementary properties of the three embedding spaces, which further validate the rationale of our proposed DECA.

2) Retrieval Results: In Fig. 6, we show retrieval results to qualitatively evaluate our proposed DECA *w.r.t.* different inference rules, as complementary to previous ablations in Tab. II and Figs. 4 and 5. Specifically, we show in each row the retrieved images using different causal effects by tweaking Eq. (7), *i.e.*, full effect, composition effect, object effect, and attribute effect, respectively from top to bottom. In general, different effects capture different visual patterns. This is especially obvious in the last two rows, *e.g.*, when retrieving *sliced fruit*, the attribute effect actually focuses on the "sliced" pattern, and it is fairly interesting to see bananas in the results, for which we deem that the model might think stacked bananas visually resembles "sliced". Again, the above results validate the effectiveness of our decomposable causal view, in both quantitative and qualitative ways.

V. CONCLUSION

In this paper, we present a novel decomposable causal view (DECA) of Compositional Zero-Shot Learning (CZSL). The core idea lies in our causal model that characterizes how compositional concepts are formed. Therein we highlight the indispensable role of contextuality between primitive concepts. To learn a causal model that recognizes this contextuality, we propose to approximate the total causal effect with three decomposable ones. An easy-to-implement pipeline is further developed to model these causal effects. We evaluate our proposed DECA on two CZSL benchmarks, showing substantial superiority over all state-of-the-art baselines. For future work, we consider probing more fine-grained causal effects

in various situations concerning compositional generalization, especially when multiple visual attributes/objects are involved.

REFERENCES

[1] Y. Atzmon, J. Berant, V. Kezami, A. Globerson, and G. Chechik, "Learning to generalize to new compositions in image understanding," *arXiv:1608.07639*, 2016. [Online]. Available: https://arxiv.org/abs/1608.07639

[2] B. M. Lake, "Towards more human-like concept learning in machines: Compositionality, causality, and learning-to-learn," Ph.D. dissertation, Massachusetts Institute of Technology, 2014.

[3] Y.-X. Wang, D. Ramanan, and M. Hebert, "Learning to model the tail," in *Proc. NeurIPS*, 2017.

[4] I. Misra, A. Gupta, and M. Hebert, "From red wine to red tomato: Composition with context," in *Proc. CVPR*, 2017, pp. 1160–1169.

[5] S. Purushwalkam, M. Nickel, A. Gupta, and M. Ranzato, "Task-driven modular networks for zero-shot compositional learning," in *Proc. ICCV*, 2019, pp. 3592–3601.

[6] T. Nagarajan and K. Grauman, "Attributes as operators: factorizing unseen attribute-object compositions," in *Proc. ECCV*, 2018, pp. 169–185.

[7] Y.-L. Li, Y. Xu, X. Mao, and C. Lu, "Symmetry and group in attribute-object compositions," in *Proc. CVPR*, 2020, pp. 11 313–11 322.

[8] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Proc. CVPR*. IEEE, 2009, pp. 1778–1785.

[9] A. Farhadi, I. Endres, and D. Hoiem, "Attribute-centric recognition for cross-category generalization," in *Proc. CVPR*. IEEE, 2010, pp. 2352–2359.

[10] A. Fathi and J. M. Rehg, "Modeling actions through state changes," in *Proc. CVPR*, 2013, pp. 2579–2586.

[11] J.-B. Alayrac, I. Laptev, J. Sivic, and S. Lacoste-Julien, "Joint discovery of object states and manipulation actions," in *Proc. ICCV*, 2017, pp. 2127–2136.

[12] J. Johnson, B. Hariharan, L. Van Der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning," in *Proc. CVPR*, 2017, pp. 2901–2910.

[13] J. Koushik, H. Hayashi, and D. S. Sachan, "Compositional reasoning for visual question answering," in *Proc. ICML*, 2017.

[14] D. A. Hudson and C. D. Manning, "Compositional attention networks for machine reasoning," in *Proc. ICLR*, 2018.

[15] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell, "Zero-shot learning with semantic output codes," in *Proc. NeurIPS*, 2009, pp. 1410–1418.

[16] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proc. CVPR*, 2009, pp. 951–958.

[17] C. H. Lampert, H. Nickisch, S. Harmeling *et al.*, "Attribute-based classification for zero-shot visual object categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 453–465, 2013.

[18] B. Romera-Paredes and P. Torr, "An embarrassingly simple approach to zero-shot learning," in *Proc. ICML*, 2015, pp. 2152–2161.

[19] S. Changpinyo, W.-L. Chao, and F. Sha, "Predicting visual exemplars of unseen classes for zero-shot learning," in *Proc. ICCV*, 2017, pp. 3476–3485.

[20] B. Chen and W. Deng, "Hybrid-attention based decoupled metric learning for zero-shot image retrieval," in *Proc. CVPR*, 2019, pp. 2750–2759.

[21] P. Zhu, H. Wang, and V. Saligrama, "Generalized zero-shot recognition based on visually semantic embedding," in *Proc. CVPR*, 2019, pp. 2995–3003.

[22] X. Zhang, S. Gui, Z. Zhu, Y. Zhao, and J. Liu, "Hierarchical prototype learning for zero-shot recognition," *IEEE Trans. Multimedia*, vol. 22, no. 7, pp. 1692–1703, 2019.

[23] S. Min, H. Yao, H. Xie, Z.-J. Zha, and Y. Zhang, "Domain-oriented semantic embedding for zero-shot learning," *IEEE Trans. Multimedia*, vol. 23, pp. 3919–3930, 2020.

[24] Z. Han, Z. Fu, S. Chen, and J. Yang, "Contrastive embedding for generalized zero-shot learning," in *Proc. CVPR*, 2021, pp. 2371–2381.

[25] E. Kodirov, T. Xiang, and S. Gong, "Semantic autoencoder for zero-shot learning," in *Proc. CVPR*, 2017, pp. 3174–3183.

[26] W. Wang, Y. Pu, V. K. Verma, K. Fan, Y. Zhang, C. Chen, P. Rai, and L. Carin, "Zero-shot learning via class-conditioned deep generative models," in *Proc. AAAI*, 2018.

[27] H. Huang, C. Wang, P. S. Yu, and C.-D. Wang, "Generative dual adversarial network for generalized zero-shot learning," in *Proc. CVPR*, 2019, pp. 801–810.

[28] J. Ni, S. Zhang, and H. Xie, "Dual adversarial semantics-consistent network for generalized zero-shot learning," in *Proc. NeurIPS*, 2019.

[29] Y. Yang, X. Zhang, M. Yang, and C. Deng, "Adaptive bias-aware feature generation for generalized zero-shot learning," *IEEE Trans. Multimedia*, 2021.

[30] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2251–2265, 2018.

[31] M. Yang, C. Deng, J. Yan, X. Liu, and D. Tao, "Learning unseen concepts via hierarchical decomposition and composition," in *Proc. CVPR*, 2020, pp. 10 245–10 253.

[32] C.-Y. Chen and K. Grauman, "Inferring analogous attributes," in *Proc. CVPR*, 2014, pp. 200–207.

[33] R. Santa Cruz, B. Fernando, A. Cherian, and S. Gould, "Neural algebra of classifiers," in *Proc. WACV*, 2018, pp. 729–737.

[34] K. Wei, M. Yang, H. Wang, C. Deng, and X. Liu, "Adversarial fine-grained composition learning for unseen attribute-object recognition," in *Proc. ICCV*, 2019, pp. 3740–3748.

[35] M. Mancini, M. F. Naeem, Y. Xian, and Z. Akata, "Open world compositional zero-shot learning," in *Proc. CVPR*, 2021, pp. 5218–5226.

[36] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, "Feature generating networks for zero-shot learning," in *Proc. CVPR*, 2018, pp. 5542–5551.

[37] Z. Nan, Y. Liu, N. Zheng, and S.-C. Zhu, "Recognizing unseen attribute-object pair with generative model," in *Proc. AAAI*, vol. 33, no. 01, 2019, pp. 8811–8818.

[38] X. Wang, F. Yu, T. Darrell, and J. E. Gonzalez, "Task-aware feature generation for zero-shot compositional learning," *arXiv:1906.04854*, 2019. [Online]. Available: https://arxiv.org/abs/1906.04854

[39] X. Wang, F. Yu, R. Wang, T. Darrell, and J. E. Gonzalez, "Tafe-net: Task-aware feature embeddings for low shot learning," in *Proc. CVPR*, 2019, pp. 1831–1840.

[40] H. Chen, Z. Nan, J. Jiang, and N. Zheng, "Learning to infer unseen attribute-object compositions," *arXiv:2010.14343*, 2020. [Online]. Available: https://arxiv.org/abs/2010.14343

[41] M. F. Naeem, Y. Xian, F. Tombari, and Z. Akata, "Learning graph embeddings for compositional zero-shot learning," in *Proc. CVPR*, 2021, pp. 953–962.

[42] Z. Xu, G. Wang, Y. Wong, and M. S. Kankanhalli, "Relation-aware compositional zero-shot learning for attribute-object pair recognition," *IEEE Trans. Multimedia*, 2021, DOI: 10.1109/TMM.2021.3104411.

[43] G. Xu, P. Kordjamshidi, and J. Y. Chai, "Zero-shot compositional concept learning," *arXiv:2107.05176*, 2021. [Online]. Available: https://arxiv.org/abs/2107.05176

[44] Y. Atzmon, F. Kreuk, U. Shalit, and G. Chechik, "A causal view of compositional zero-shot recognition," in *Proc. NeurIPS*, vol. 33, 2020, pp. 1462–1473.

[45] F. Ruis, G. Burghours, and D. Bucur, "Independent prototype propagation for zero-shot compositionality," in *Proc. NeurIPS*, 2021.

[46] P. Judea, "Causality: models, reasoning, and inference," *Cambridge University Press*, vol. 521, no. 77362, p. 8, 2000.

[47] J. Pearl, "Direct and indirect effects," in *Proc. UAI*, 2001, p. 411–420.

[48] J. Pearl and D. Mackenzie, *The book of why: the new science of cause and effect*. Basic Books, 2018.

[49] J. M. Robins, "Semantics of causal dag models and the identification of direct and indirect effects," *Oxford Statistical Science Series*, pp. 70–82, 2003.

[50] D. B. Rubin, "Causal inference using potential outcomes: Design, modeling, decisions," *Journal of the American Statistical Association*, vol. 100, no. 469, pp. 322–331, 2005.

[51] S. L. Morgan and C. Winship, *Counterfactuals and causal inference*. Cambridge University Press, 2015.

[52] B. Schölkopf, "Causality for machine learning," *arXiv:1911.10500*, 2019. [Online]. Available: https://arxiv.org/abs/1911.10500

[53] W. Qin, H. Zhang, R. Hong, E.-P. Lim, and Q. Sun, "Causal interventional training for image recognition," *IEEE Trans. Multimedia*, 2021.

[54] T. Wang, C. Zhou, Q. Sun, and H. Zhang, "Causal attention for unbiased visual recognition," in *Proc. ICCV*, 2021, pp. 3091–3100.

[55] Y. Niu, K. Tang, H. Zhang, Z. Lu, X.-S. Hua, and J.-R. Wen, "Counterfactual vqa: A cause-effect look at language bias," in *Proc. CVPR*, 2021, pp. 12 695–12 705.

[56] G. Nan, R. Qiao, Y. Xiao, J. Liu, S. Leng, H. Zhang, and W. Lu, "Interventional video grounding with dual contrastive learning," in *Proc. CVPR*, 2021, pp. 2764–2774.

[57] K. Tang, J. Huang, and H. Zhang, "Long-tailed classification by keeping the good and removing the bad momentum causal effect," in *Proc. NeurIPS*, vol. 33, 2020, pp. 1513–1524.

[58] Z. Yue, H. Zhang, Q. Sun, and X.-S. Hua, "Interventional few-shot learning," in *Proc. NeurIPS*, vol. 33, 2020, pp. 2734–2746.

[59] Z. Yue, T. Wang, Q. Sun, X.-S. Hua, and H. Zhang, "Counterfactual zero-shot and open-set visual recognition," in *Proc. CVPR*, 2021, pp. 15 399–15 409.

[60] J. Mitrovic, B. McWilliams, J. Walker, L. Buesing, and C. Blundell, "Representation learning via invariant causal mechanisms," in *Proc. ICLR*, 2021.

[61] W.-L. Chao, S. Changpinyo, B. Gong, and F. Sha, "An empirical study and analysis of generalized zero-shot learning for object recognition in the wild," in *Proc. ECCV*, 2016, pp. 52–68.

[62] P. Isola, J. J. Lim, and E. H. Adelson, "Discovering states and transformations in image collections," in *Proc. CVPR*, 2015, pp. 1383–1391.

[63] A. Yu and K. Grauman, "Fine-grained visual comparisons with local learning," in *Proc. CVPR*, 2014, pp. 192–199.

[64] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv:1807.03748*, 2018. [Online]. Available: https://arxiv.org/abs/1807.03748

[65] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. CVPR*, 2020, pp. 9726–9735.

[66] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar *et al.*, "Bootstrap your own latent: A new approach to self-supervised learning," in *Proc. NeurIPS*, 2020.

[67] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. ICML*, 2020.

[68] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Proc. NeurIPS*, 2020.

[69] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proc. CVPR*, 2021, pp. 15 750–15 758.

[70] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv:1503.02531*, 2015. [Online]. Available: https://arxiv.org/abs/1503.02531

[71] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.

[72] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[73] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv:1607.06450*, 2016. [Online]. Available: https://arxiv.org/abs/1607.06450

[74] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proc. ICML*, 2010.

[75] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[76] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. NeurIPS*, 2013.

[77] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguist.*, vol. 5, pp. 135–146, 2017.

[78] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[79] L. Van Der Maaten, "Accelerating t-sne using tree-based algorithms," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3221–3245, 2014.

[80] S. Gidaris and N. Komodakis, "Dynamic few-shot visual learning without forgetting," in *Proc. CVPR*, 2018, pp. 4367–4375.

[81] B. Liu, Y. Cao, Y. Lin, Q. Li, Z. Zhang, M. Long, and H. Hu, "Negative margin matters: Understanding margin in few-shot classification," in *Proc. ECCV*, 2020.

[82] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. EMNLP*, 2014, pp. 1532–1543.

Muli Yang received the B.E. degree from Xidian University, Xi'an, China, in 2017, where he is currently pursuing the Ph.D. degree with the School of Electronic Engineering. His research interests include multimodal reasoning, causal inference, computer vision, and machine learning.

Chenghao Xu received the B.E. degree from Xidian University, Xi'an, China, in 2020, where he is currently pursuing the Ph.D. degree with the School of Electronic Engineering. His research interests include multimodal analysis, computer vision and machine learning.

Aming Wu received the Ph.D. degree from Tianjin University, Tianjin, China, in 2021. He joined Xidian University as a pre-tenured associate professor at the school of electronic engineering in Jan. 2021. His current research interests include computer vision, multimedia analysis, and machine learning.

Cheng Deng (Senior Member, IEEE) received the B.E., M.S., and Ph.D. degrees in signal and information processing from Xidian University, Xi'an, China. He is currently a Full Professor with the School of Electronic Engineering, Xidian University. He is the author and the coauthor of more than 100 scientific articles at top venues, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CYBERNETICS, NeurIPS, ICML, CVPR, ICCV, AAAI, IJCAI, and KDD. His research interests include computer vision, pattern recognition, and information hiding.