Learning Invariant Visual Representations for Compositional Zero-Shot Learning

Tian Zhang[1*], Kongming Liang[1*], Ruoyi Du[1], Xian Sun[2], Zhanyu Ma[1], and Jun Guo[1]

[1] Pattern Recognition and Intelligent System Laboratory, School of Artificial Intelligence, Beijing University of Posts and Telecommunications
{zhangtian1874,liangkongming,duruoyi,mazhanyu,guojun}@bupt.edu.cn
[2] Aerospace Information Research Institute, Chinese Academy of Sciences
sunxian@aircas.ac.cn

Abstract. Compositional Zero-Shot Learning (CZSL) aims to recognize novel compositions using knowledge learned from seen attribute-object compositions in the training set. Previous works mainly project an image and a composition into a common embedding space to measure their compatibility score. However, both attributes and objects share the visual representations learned above, leading the model to exploit spurious correlations and bias towards seen pairs. Instead, we reconsider CZSL as an out-of-distribution generalization problem. If an object is treated as a domain, we can learn object-invariant features to recognize the attributes attached to any object reliably. Similarly, attribute-invariant features can also be learned when recognizing the objects with attributes as domains. Specifically, we propose an invariant feature learning framework to align different domains at the representation and gradient levels to capture the intrinsic characteristics associated with the tasks. Experiments on three CZSL benchmarks demonstrate that the proposed method significantly outperforms the previous state-of-the-art.

Keywords: Compositional Zero-Shot Learning, Out-of-Distribution Generalization, Invariant Feature Learning

1 Introduction

Humans can easily generalize the *red* state from *apples* to *tomatoes* even if no images of *red tomatoes* have been seen. Since visual concepts follow the long tailed distribution, the instances of most concepts are rarely presented in the real world scenario. Therefore, the ability to generalize the learned knowledge to novel concepts is of vital importance for human to recognize a large number of concepts and is considered as one of the hallmarks of human intelligence [22,29]. The goal of Compositional Zero-Shot Learning (CZSL) is to build a model that can learn the attributes and objects from seen compositions and generalize them well to unseen compositions. For instance, the model trained with images of *red apples* and *green tomatoes* can correctly predict images of *red tomatoes*.

[*] Equal contribution; codes are available at https://github.com/PRIS-CV/IVR.

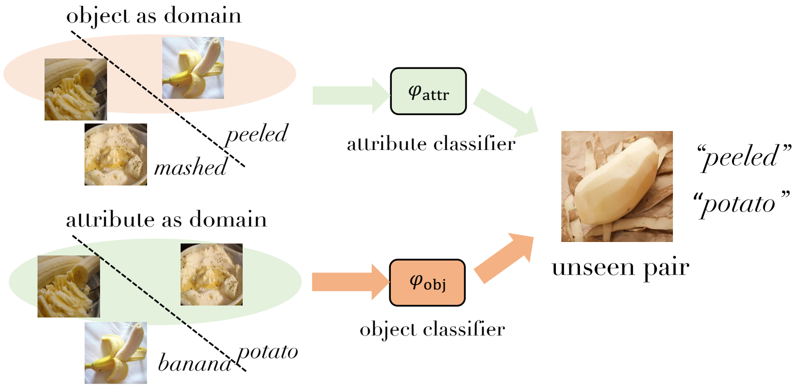

Fig. 1. The illustration of our motivation. Ellipses represent corresponding domains. The samples in one ellipse belong to the same domain, the samples outside the ellipse belong to other domains. And the dotted lines represent category decision boundaries within decoupled feature space.

Previous works [22,24,19,21] in CZSL mainly project image features and attribute-object composition features into a common embedding space and constrain the features belonging to the same concept to be closer. Specifically, the current state-of-the-art method [21] use cosine similarity to calculate the compatibility score of images and compositions in the embedding space. Since the features are learned in a composition way, they are not disentangled for attribute and object which makes the model over-rely on a limited number of attribute-object pairs in the training process. For instance, when machines had only seen *red apples*, they might easily misidentify *red tomatoes* as *red apples* since classifier had prone to spuriously correlate *red* with *apple*. Machine learning models are data-driven and typically require samples of various perspectives and lighting. This makes them often rely on spurious features [2,33,43,14,1] unrelated to the core concept and lose generalization performance [10], especially in zero-shot learning scenarios. Therefore, recognizing attributes and objects independently may actually assist the model in achieving better performance.

In this paper, we leverage the idea of Domain Generalization (DG) to improve the ability of the model to generalize to unknown compositions. Most deep learning methods work well under the *i.i.d.* assumption that training and testing data are independently and identically distributed [27,4]. However, this assumption does not always hold true in reality. When the probability distributions of training and testing data are different, the performance of deep learning models is often degraded due to the domain shift [31,38]. DG trains model only with data from the source domain, making it generalize well to the unseen arbitrary target domain. For instance, given a training set consisting of *photos*, *cartoon images* and *paintings*, DG requires training a model to have promising performance in classifying *sketches*, which are significantly different from the images in the training set. Most of the work alleviates domain shift by aligning feature distributions of the source with target domains, resulting in domain-invariant features.

Since a domain is composed of data that are sampled from a distribution [38], the Compositional Zero-Shot Learning task is analogous to two DG sub-tasks in essence, by taking objects or attributes as domains. As shown in the Figure 1, in the case of treating objects as domains, if the model learns the attributes of *mashed* and *peeled* in the *banana* domain, then we expect that it can also reliably recognize the attribute of *peeled* when generalized to the *potato* domain. Similarly, in the case of treating attributes as domains, if the model learns the objects of *banana* and *potato* in the *mashed* domain, it should recognize the object of *potato* when generalized to the *peeled* domain. Eventually, the model is able to successfully recognize the unseen pairs (*peeled potato*). We simulate a domain generalization scenario by designing a triplet input network. To decouple the highly-coupled features, we construct two branches, the object-domain branch and the attribute-domain branch. For the object-domain branch, our goal is to accurately recognize the attribute regardless of object labels. We learn consistency at the representation level by discarding object-specific channels. Moreover, we minimize the gradient differences of attribute prediction in different object domains to achieve gradient-level consistency. For the attribute-domain branch, we learn attribute-invariant features in the same way. Finally, by penalizing domain-specific power of features, we discover invariant mechanisms in the data which are hard to vary across examples and thus learn the optimal attribute classifier and object classifier.

The contributions of the paper are summarized below. (1) To the best of our knowledge, we are the first to solve the Compositional Zero-Shot Learning task from a Domain Generalization perspective. In other words, the compositional learning problem is transformed into a domain-shift problem. (2) We treat attributes or objects as domains and align different domains to learn domain-invariant features, thus improving the generalization performance of the model to recognize unseen pairs. (3) We prove the effectiveness of our method through abundant experiments.

2 Related works

2.1 Compositional Zero-Shot Learning

Compositional Zero-Shot Learning (CZSL) is a special case of Zero-Shot Learning (ZSL) [25,41,40]. Given a training set containing a set of attribute-object compositions, CZSL aims to recognize unknown compositions of these attributes and objects at inference time. Part of the work proposes to learn classifiers for individual concepts and combine them to recognize integrated concepts. Chen et al. [8] deduce unobserved attribute-object pairs through tensor decomposition during training. Misra et al. [22] consider compositionality and contextuality as the key to solving CZSL, and they merge classifiers for primitive concepts into classifiers for composite concepts. A most popular line of work involves embedding attribute-object compositions into a feature space. Nagarajan et al. [24]

argue that objects are entities while attributes are properties of the objects and consider the composition of attributes and objects as a learned transformation. Wei et al. [39] model the attribute-object relationships within the feature space based on a GAN framework. Li et al. [19] propose symmetry as an essential principle for attribute-object transformations and introduce group theory as an axiomatic foundation to satisfy the specific principles of nature. Mancini et al. [21] propose a new open world setting for CZSL task where the prior knowledge of unseen compositions is not provided. Instead, other works learn the joint compatibility between the input image and the attribute-object pair. Purushwalkam et al. [29] train a set of network modules jointly with a gating network to produce features that indicate compatibility between the input image and the concept. Atzmon et al. [3] describe CZSL from a causal perspective and try to find which intervention cause the image. Unlike these works, we focus on the independence between the sub-concepts and learn an attribute classifier and object classifier that can be generalized to new compositions.

2.2 Domain Generalization

In reality, the distribution of training and test sets is often different, leading to model performance degradation. This problem is known as out-of-distribution generalization or domain generalization [5,23,45]. Since the generalization ability of the model often depends on the quantity and quality of training data [38], one line of work increases the diversity of existing training data through data augmentation and data generation to learn more general representations. Qiao et al. [30] leverage Wasserstein Auto-Encoders (WAE) [37] to help generate samples that retain semantics and have large domain transportation. Shankar et al. [34] introduce a domain classifier to expand the training data by disturbing the input data. Carlucci et al. [6] enrich the understanding of the data by solving puzzle problems, allowing the model to induce invariance and regularity autonomously. A different line of work uses domain alignment techniques or feature disentanglement to learn domain-invariant features. Sun et al. [36] conduct domain alignment by matching the mean and variance of representations in different domains. Li et al. [18] use Maximum Mean Discrepancy (MMD) to align different domains to obtain domain-invariant representation. Peng et al. [27] decouple features into domain-invariant features, domain-specific features, and class-irrelevant features through adversarial learning. Huang et al. [12] propose a self-challenge mechanism, which iteratively discards the dominant features activated on the training data. Kim et al. [15] propose self-supervised contrastive regularization to map the latent representations of the positive pair samples close together. In this paper, we mainly leverage the idea of exploring invariance in DG to enhance the performance of the CZSL task.

3 Methods

In typical CZSL, we have access to all the attributes and objects, while only part of attribute-object compositions can be obtained in the training phase. The

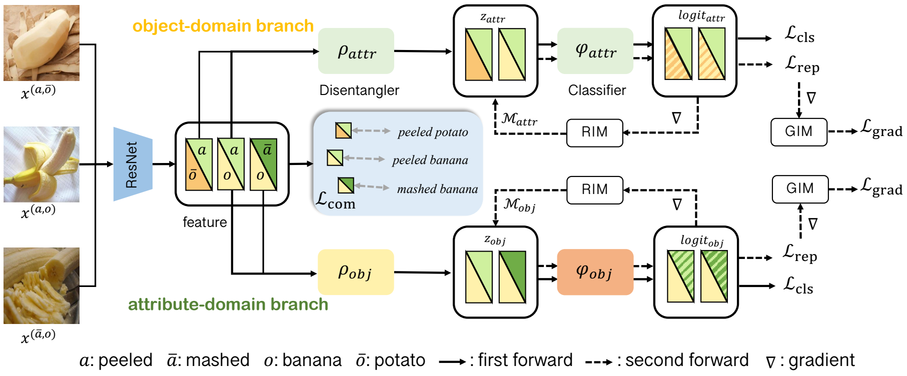

Fig. 2. Overview of the proposed framework. We construct object-domain branch and attribute-domain branch. In the object-domain branch, we execute consistent alignment across different object domains so that the model learns the essential characteristics of the attribute. The same for the attribute-domain branch.

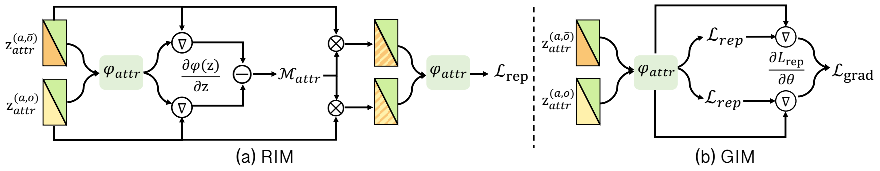

Fig. 3. Consider the object-domain branch with the representations $z^{(a,o)}_{attr}$ and $x^{(a,\bar{o})}_{attr}$. (a) In the representation invariant mechanism (RIM), we learn object-irrelevant attribute features at the representation level by filtering out object-specific channels. (b) In the gradient invariant mechanism (GIM), we learn object-irrelevant attribute features at the gradient level by minimizing the distance between gradients of different object domains.

goal is to recognize unknown compositions of individual attribute and object concepts. Composing the learned knowledge into unseen compositions heavily relies on out-of-distribution generalization ability [16,17,33,14,1]. Therefore, we formalize the CZSL problem into two domain generalization sub-tasks, in which we consider attributes as domains to recognize objects and vice versa. Then, two types of invariant mechanisms are proposed to remove the spurious domain-specific features and improve the generalization ability of the model.

An overview of our proposed framework is shown in Figure 2. In the following sections, we first introduce the visual and composition embedding learning procedure. Then we present how the visual features are decomposed and processed by representation and gradient invariant mechanisms in sequential. Finally, we describe the training and inference methodologies.

3.1 Visual and Composition Embedding

We need to train a model that learns a mapping from a set of images X to a set of compositions $Y = Y_{attr} \times Y_{obj}$, where Y_{attr} is a set of attribute labels and Y_{obj} is a set of object labels. The composition label is divided into $Y = Y_s \cup Y_u$, where Y_s is the set of seen compositions during training and Y_u is the set of unseen compositions for the validation and test sets, with $Y_s \cap Y_u = \emptyset$. Given an image $x \in X$ in the training set and its corresponding label $y \in Y_s$, we first use a pre-trained network $f(\cdot)$ (e.g., ResNet-18 [11]) to extract its visual embedding. Then, the composition embedding function $g(\cdot)$ projects the combined concepts y into a common semantic space. The composition classification loss can be obtained by minimizing the distance between the two embedding features,

$$h_{comp}(x, y) = d_{cos}(f(x), g(y)), \tag{1}$$

where $d_{cos}(\cdot, \cdot)$ is the cosine distance of the input two embeddings. The distance in the embedding space represents the compatibility of the input image and the attribute-object composition. Therefore, the smaller the distance is, the higher probability that the composition exists in the image [24].

However, the visual representations learned in the above manner are shared by both attributes and objects, which may lead the model exploiting spurious correlations and bias the model against seen pairs. In this work, we utilize invariant feature learning to decouple attributes and objects from a domain generalization perspective. The learned invariant features explore the independence between attribute and object concepts and prove to be effective to complement the conventional visual embedding.

3.2 Decomposing Visual Features

In order to conduct the invariant feature learning for CZSL, we need to decompose the visual features into two parts by considering object and attribute as domain respectively. Here, we design a triplet input network with $x^{a,\bar{o}}$, $x^{a,o}$ and $x^{\bar{a},o}$ as inputs to diversify the inter-domain variation, where $a, \bar{a} \in Y_{attr}$ denote different attributes, and $o, \bar{o} \in Y_{obj}$ denote different objects, e.g., $x^{(a,\bar{o})}$ represents an image of different object with the same attribute as $x^{(a,o)}$. We denote the composition set of a triplet input as $\mathcal{C} = \{(a, \bar{o}), (a, o), (\bar{a}, o)\}$. And the classification task set is denoted as $\mathcal{T} = \{attr, obj\}$.

The extracted visual features from the pre-trained network $f(\cdot)$ are directly fed into two individual MLPs, attribute disentangler $\rho_{attr}(\cdot)$ and object disentangler $\rho_{obj}(\cdot)$. For $i \in \mathcal{C}$ and $j \in \mathcal{T}$, the image features of x^i can be decoupled as $z^i_j = \rho_j(f(x^i))$. Given the cross entropy loss function $l(\cdot, \cdot)$, the attribute and object classification loss can be defined as,

$$L_{cls} = \sum_{i \in \mathcal{C}} \sum_{j \in \mathcal{T}} l(\varphi_j(z^i_j; \theta_j), y^i_j), \tag{2}$$

where $\varphi_j(\cdot)$ denotes the classifier of task j which predict classification labels over the decomposed visual features. θ_j represents the parameters of classifier $\varphi_j(\cdot)$.

3.3 Learning Invariant Features for CZSL

A notion of invariance implies something that stays the same while something else changes [26]. Capturing invariance helps model learn the core features related to the label. Returning to the previous example, the explanations to distinguish *tomatoes* from *apples* should be invariant, no matter whether the *tomatoes* are *red* or *green*. Therefore, we leverage invariant feature learning to capture the invariance of objects when attributes change or vice versa. Finally the learned invariant features of attributes and objects can be generalized to novel compositions.

When the model takes $x^{(a,o)}$ and $x^{(a,\bar{o})}$ as inputs, we construct a scenario that recognize attribute concept with object as domain. Similarly, with $x^{(a,o)}$ and $x^{(\bar{a},o)}$ as inputs, we construct a scenario that recognize object concept in terms of attribute as domain. Our goal is to recognize an attribute associated with any objects and recognize an object described by any attributes. To improve the generalization performance of the model, we explicitly promote invariance to disentangle spurious features at representation level and gradient level.

Representation Invariant Learning. To learn an invariant classifier that helps with generalizing to new domain, we explore invariance at the representation level to pull together samples with the same class from different domains in the feature space. In other words, learn a model that maps different domains to a single statistical distribution [2,33].

Firstly, we calculate the gradient of prediction results over the different domains with respect to the representation,

$$g_j^i = \frac{\partial([\varphi_j(z_j^i;\theta_j)]^\top \cdot y_j^i)}{\partial z_j^i}. \tag{3}$$

The representations associated with the similar gradients indicate intrinsic characteristic of attribute concepts that are invariant to object factors or vice versa. Thus we calculate the absolute value of the difference between the two gradients,

$$\varDelta g_{attr} = \left| g_{attr}^{(a,o)} - g_{attr}^{(a,\bar{o})} \right|, \ \varDelta g_{obj} = \left| g_{obj}^{(a,o)} - g_{obj}^{(\bar{a},o)} \right|. \tag{4}$$

The semantic channels with small difference can be regarded as object-invariant feature channels of attribute and attribute-invariant feature channels of object. We sort the difference from the largest to the smallest, taking the value at α percent, and denoted as t^α. Then we construct a mask that shares the same dimension with the representation as follows. For the k^{th} element,

$$m_j(k) = \begin{cases} 0, & \text{if } \varDelta g_j(k) \geq t^\alpha \\ 1, & \text{else} \end{cases}. \tag{5}$$

By overwriting the mask to the original representation, the network filters out domain-specific feature channels to learn the domain-invariant feature,

$$\hat{z}_j^i = z_j^i \odot m_j. \tag{6}$$

Then we computes the cross entropy loss with the object-irrelevant attribute-specific representation and the attribute-irrelevant object-specific representation,

$$\begin{aligned} L_{rep} = l(\varphi_{attr}(\hat{z}_{attr}^{(a,o)};\theta_{attr}), y_{attr}^{(a,o)}) + l(\varphi_{attr}(\hat{z}_{attr}^{(a,\bar{o})};\theta_{attr}), y_{attr}^{(a,\bar{o})}) \\ + l(\varphi_{obj}(\hat{z}_{obj}^{(a,o)};\theta_{obj}), y_{obj}^{(a,o)}) + l(\varphi_{obj}(\hat{z}_{obj}^{(\bar{a},o)};\theta_{obj}), y_{obj}^{(\bar{a},o)}). \end{aligned} \tag{7}$$

Gradient Invariant Learning. Since reducing empirical risk [44] across different domains can reduce the sensitivity of models to distribution shift [17], we execute gradient-level domain alignment to optimize different domains in the same direction, which will penalize the network to minimize the dispersion of gradients in different domains to capture invariance. The objective of enhancing gradient consistency is to find local or global minimum in the loss space across all of the training domains and let the network share similar Hessians for different domains [33].

We calculate the gradient of attribute prediction results to attribute classifier in different object domains as well as the gradient of object prediction results to object classifier in different attribute domains as follows,

$$G_j^i = \frac{\partial l(\varphi_j(\hat{z}_j^i;\theta_j), y_j^i)}{\partial \theta_j}. \tag{8}$$

The gradient represents the optimal path. It is easier to obtain invariant predictions in different domains by encouraging the same optimization paths in all domains [35]. In order to align different domains at the gradient level and learn the invariance associated with label, we penalize the domain prediction ability by minimizing the Euclidean distance $d_{euc}(\cdot,\cdot)$ between the two gradients as shown below,

$$L_{grad} = d_{euc}(G_{attr}^{(a,o)}, G_{attr}^{(a,\bar{o})}) + d_{euc}(G_{obj}^{(a,o)}, G_{obj}^{(\bar{a},o)}). \tag{9}$$

We measure the alignment by calculating the Euclidean distances of gradients across different domains. In addition, cosine distance is also considered to measure the alignment of domains in the ablation experiment (see Section **4.6**).

By introducing these regularizing terms, we can adaptively look for domain-specific channels and discard them, forcing the network to find an invariant relationship between the input image and the label at the representation-level consistency. We also conduct all the domains optimized in the same direction at the gradient-level consistency. Finally, we get decoupled attribute features and object features, which will improve the predictive performance of the model in unseen compositions.

3.4 Training and Inference

For training, we borrow from previous works using the composition classification loss in embedding learning to explore the dependence between attributes and objects,

$$L_{comp} = \sum_{i \in \mathcal{C}} h_{comp}(x^i, y^i). \tag{10}$$

Simultaneously, we employ invariant feature learning to decouple attributes and objects to explore their independence. Finally, the objective of optimization can be expressed as,

$$L = L_{comp} + L_{cls} + \lambda_1 L_{rep} + \lambda_2 L_{grad}, \tag{11}$$

where λ_1 and λ_2 are trade-off parameters.

During inference, given an image in the test set, we project it into the common embedding space. The distance between visual embedding features and all candidate pair vectors is calculated and sorted to obtain a pair score predicted in the form of coupling. On the other hand, we use classifiers to predict attributes and objects separately in a decoupled manner and combine the predicted results into a pair score. The final prediction result is obtained by adding the two pair scores, which will improve the performance of the model for both seen and unseen pairs.

4 Experiments

4.1 Datasets

Mit-States [13] and UT-Zappos50K [42] are two benchmark datasets widely used in CZSL task.

After careful observation of the dataset, we also discover three significant problems. First, because Mit-States is labelled automatically using early image search engine technology [3], it contains much noise. For example, there is an image labelled *pierced bear*, but it is actually a brown ceramic pot. Second, the existence of both super-classes and sub-classes in this dataset, such as *animal* and *horse*, as well as *fruit* and *apple*, can create ambiguity. Thirdly, the semantic expression of some attributes is not clear enough. For example, *big bear* and *large bear* are precisely the same from the picture. In light of these issues, we believe that the Mit-States dataset is too noisy to evaluate effectively. Therefore, we use UT-Zappos50K, Clothing16K, and AO-CLEVr for the experiment.

UT-Zappos50K [42] is a fine-grained shoes dataset which contains about 33k images with 12 object classes and 16 attribute classes. The object concepts are mainly the types of shoes (e.g. heels, slippers), while the attribute concepts are mainly the material of shoes (e.g. canvas, leather). Following the generalized evaluation protocol proposed by [29], we test on both seen and unseen pairs. We adopt the standard split from [29,21], the training set has about 23k images belonging to 83 attribute-object pairs. The validation set has about 3k images consisting of 15 seen pairs and 15 unseen pairs. And the test set has about 3k images consisting of 18 seen pairs and 18 unseen pairs.

Clothing16K[3] was initially a dataset used for multi-label classification in Kaggle competitions with 8 object classes and 9 attribute classes. The object concepts are mainly the types of clothing (e.g. shirt, pants), while the attribute

[3] `https://www.kaggle.com/kaiska/apparel-dataset`

concepts are mainly the clothing colour(e.g. black, green). We find that the attributes and objects of this dataset are very distinct and almost contain no noise, which is very suitable for the CZSL task. Therefore, we split the dataset by ourselves following the generalized ZSL principle [29]. The training set has about 7k images in 18 attribute-object pairs. The validation set consists of 10 seen pairs and 10 unseen pairs with a total of about 5k images. And the test set consists of 9 seen pairs and 8 unseen pairs with a total of about 3k images.

AO-CLEVr [3] is a synthetic dataset consisting of 3 object classes (e.g. sphere, cube, cylinder) and 8 attribute classes (e.g. yellow, gray), with 24 compositional classes in total. We also split the dataset following the generalized ZSL principle [29]. The training set has about 103k images in 16 attribute-object pairs. The validation set consists of 4 seen pairs and 4 unseen pairs with a total of about 39k images. And the test set has about 38k images from 4 seen pairs and 4 unseen pairs.

4.2 Metric

Following [29,21], we test the performance by the accuracy of their top-1 prediction for recognizing seen pairs (*Seen*) and unseen pairs (*Unseen*) in the validation set and test set. To account for the inherent bias towards seen pairs, we follow Chao et al. [7] to add a calibration bias term to the unseen pairs to balance the seen-unseen accuracy. When the calibration value is positive, the prediction accuracy of the unseen pair will be high, and when the calibration value is negative, the model tends to have a bias towards seen pairs. As the candidate value changes, a curve can be drawn with the accuracy of seen pairs on the X-axis and unseen pairs on the Y-axis. We report the Area Under Curve (AUC) to evaluate the overall performance. We also consider the best harmonic mean (HM) of seen accuracy and unseen accuracy defined as $2(Seen * Unseen)/(Seen + Unseen)$ in this curve, which can penalize the large performance discrepancies between two quantities and as such enables the model to verify performance on both seen and unseen pairs simultaneously.

4.3 Implementation Details

Following [29,21], we use ResNet-18 [11] pretrained on ImageNet [9] as the feature extractor. For a fair comparison with prior works, we do not finetune this network. The extracted 512-dimension features are mapped into a common embedding space through an image embedding function consists of 2 fully-connected layers. Then, we build an attribute disentangler and an object disentangler with a fully-connected layer to map the features into attribute subspace and object subspace respectively. Finally, an attribute classifier and an object classifier implemented by a fully-connected layer are trained to recognize concepts respectively. Simultaneously, we map concatenated compositional text features into the common embedded space. We use Adam optimizer with a initial learning rate set to 0.001 and a weight decay set to 5×10^{-5}. The λ_1 and λ_2 in Eqs. (11) are respectively set to 1 and 10 in all experiments.

4.4 Compared Methods

We compare our work with several methods.

(1) LE+ [24] uses GloVe [28] word vectors to represent attribute and object concepts and trains the neural network to project the concatenated concept features and visual features to a joint embedding space.

(2) AttrAsOp [24] treats the attribute as a matrix operator and treats the object as a vector. Then conducts attribute-conditioned transformations to learn unseen attribute-object pairs.

(3) SymNet [19] considers the symmetry principle in the attribute-object composition process and introduces group theory as a foundation for axiomatics.

(4) TMN [29] trains a set of network modules jointly with a gating network where the compositional reasoning task is divided into sub-tasks that multiple small networks can solve in a semantic concept space.

(5) CompCos [21] proposes an open world setting where all the compositions of attributes and objects could potentially exist. A feasible strategy is proposed to remove the impossible compositions.

(6) VisProd [20]. Unlike the above methods, VisProd does not model the composition explicitly but imposes attribute classifier and object classifier independently over the image features. The prediction result of a composition is the product of the probability of each element: $P(c) = P(a) \times P(o)$.

4.5 Quantitative Result

We summarize the results for our method and other methods on the three datasets in Table 1. Our method outperforms almost all reported results. Compared with the accuracy of seen pairs, our method improves the accuracy of unseen pairs to a greater extent. This is because our method inevitably loses the spurious correlation between attributes and objects while learning them independently. In other words, it hurts the model's bias against the seen pairs. Although the ability of model to recognize seen pairs is weak, HM and AUC, the metrics of comprehensive recognition ability, increased. The experimental result sufficiently proves the superiority of our proposed method.

Figure 4 shows the unseen-seen accuracy curve on the UT-Zappos50K and Clothing16K dataset. With the increase of calibration value, the classification accuracy of seen pairs decreases while that of unseen pairs increases. This is a general and essential trade-off when learning models that are robust for interventions [32]. Compared to other methods, our method keeps a better balance between seen and unseen pairs on both datasets, which leads to better performance.

Overall, the results on these challenging datasets strongly support our idea of leveraging invariant mechanisms to decouple attributes and objects effectively. Learning attributes and objects in a decoupled way may discourage certain types of correlations [3], so the model can not benefit from them when the test and training distributions are the same, that is, recognizing seen pairs. However, when recognizing unseen pairs, where the test and training distributions are

Table 1. Comparative experiment between recent methods with our method on UT-Zappos50K, Clothing16K, and AO-CLEVr.

	UT-Zappos50K				Clothing16K				AO-CLEVr			
Method	Seen	Unseen	HM	AUC	Seen	Unseen	HM	AUC	Seen	Unseen	HM	AUC
LE+ [24]	53.0	61.9	41.0	25.7	93.9	88.3	77.4	76.0	95.7	99.2	92.3	93.5
AttrAsOp [24]	59.8	54.2	40.8	25.9	95.1	80.1	60.0	58.7	95.5	85.5	64.8	65.8
SymNet [19]	49.8	57.4	40.4	23.4	95.7	90.2	73.4	75.2	87.1	97.8	71.8	74.2
VisProd [20]	56.6	60.2	43.7	28.1	96.4	91.4	74.7	77.5	91.9	98.2	71.3	75.6
TMN [29]	58.7	60.0	45.0	29.3	94.9	89.7	80.9	79.5	96.1	93.9	86.9	87.1
CompCos [21]	**59.8**	62.5	43.1	28.7	96.9	93.0	83.9	84.7	96.3	99.1	94.5	94.2
Ours	56.9	**65.5**	**46.2**	**30.6**	**96.9**	**94.6**	**86.3**	**87.0**	**97.1**	**99.3**	**95.1**	**95.6**

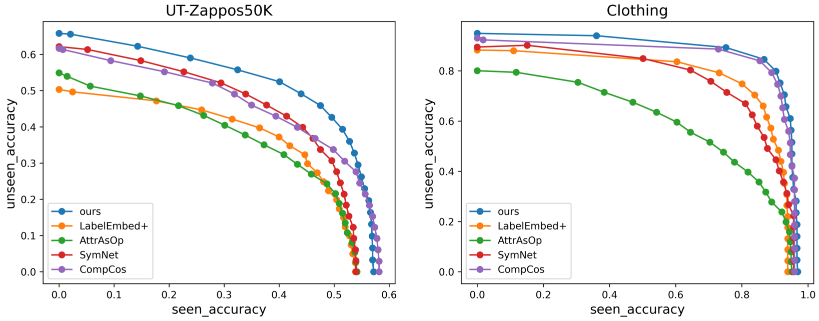

Fig. 4. Unseen-seen accuracy on UT-Zappos50K and Clothing16K under various calibration biases.

different, our method of improving generalization performance can come into play without taking advantage of these spurious correlations.

4.6 Ablation Study

To verify the effect of each proposed component, we conduct ablation experiments on the UT-Zappos50K and Clothing16K datasets. As shown in Table 2, when only compositional classification loss (denoted as "L_{comp}") is applied, the model have a positive bias towards the seen pairs because of the dependence between objects and attributes. When the concepts are learned in a decoupled way using attribute and object classifiers (denoted as "L_{cls}"), the model is biased towards unseen pairs since the correlation between attributes and objects is removed. The utilization of representation invariant mechanism (denoted as "L_{rep}") can help the model to discard domain-specific spurious features at the representation level, thus improving the performance of the model. When the gradient invariant mechanism (denoted as "L_{grad}") is employed, the gradients of different domains are optimized in the same direction. Through these two invariant learning mechanisms, the model can learn the optimal attribute classifier

Table 2. Analysis of each component on UT-Zappos50K and Clothing16K.

	UT-Zappos50K				Clothing16K			
Method	Seen	Unseen	HM	AUC	Seen	Unseen	HM	AUC
L_{comp}	**58.4**	58.0	43.5	27.8	96.9	91.8	81.6	82.9
L_{cls}	56.0	63.5	44.0	27.7	95.1	93.5	82.2	83.2
$L_{cls}+L_{comp}$	57.0	63.4	44.2	28.8	96.2	93.7	84.7	84.8
$L_{cls}+L_{comp}+L_{rep}$	55.9	65.5	45.3	29.8	96.7	94.0	85.3	85.3
$L_{cls}+L_{comp}+L_{grad}$	56.6	64.4	46.1	30.0	**97.2**	94.2	85.6	86.3
$L_{cls}+L_{comp}+L_{rep}+L_{grad}$	56.9	**65.5**	**46.2**	**30.6**	96.9	**94.6**	**86.3**	**87.0**

Table 3. Analysis of parameter α on UT-Zappos50K and Clothing16K.

	UT-Zappos50K				Clothing16K			
α	Seen	Unseen	HM	AUC	Seen	Unseen	HM	AUC
1/6	56.9	**65.5**	**46.2**	**30.6**	96.8	94.4	86.4	86.6
1/4	56.3	65.0	45.5	29.9	**96.9**	**94.6**	86.3	**87.0**
1/3	**57.5**	63.4	45.2	29.2	96.8	94.5	86.4	86.6
1/2	53.6	65.4	44.4	28.5	96.8	93.9	**86.7**	86.5

and object classifier, which remarkably improves the comprehensive performance of the model.

Effect of parameter α. The scale parameter α is employed to control the proportion of discarding in Eqs. (5). We select α in $\left\{\frac{1}{6}, \frac{1}{4}, \frac{1}{3}, \frac{1}{2}\right\}$ and report the performance of the model in Table 3. For the UT-Zappos50K dataset, the optimal performances can be observed when α is set to $\frac{1}{6}$. For the Clothing16K dataset, the optimal performances can be observed when α is set to $\frac{1}{4}$. A suitable α can subtly discard domain-specific features and help the model generalize from known concepts to unseen ones by using domain-invariant features.

Effect of distance function. In the gradient invariant mechanism, we use Euclidean distance to measure the distance between gradients in different domains. In addition, our method also works with cosine distance. As shown in Table 4, the performance of Euclidean distance is better than cosine distance, probably because we pay more attention to the absolute numerical differences between gradients.

4.7 Image Retrieval

To qualitatively evaluate our method, we further report image retrieval results. Figure 5 shows examples of retrieving images. The query is made up of attribute text and object text. We choose compositions of different objects with the same attribute and compositions of different attributes with the same object. For UT-Zappos50K and Clothing16K datasets, our method can retrieve a certain

Table 4. Analysis of distance function on UT-Zappos50K and Clothing16K.

	UT-Zappos50K				Clothing16K			
Distance Function	Seen	Unseen	HM	AUC	Seen	Unseen	HM	AUC
Euclidean	56.9	**65.5**	**46.2**	**30.6**	**96.9**	**94.6**	**86.3**	**87.0**
Cosine	**58.3**	62.6	44.2	28.6	96.7	94.3	85.0	85.9

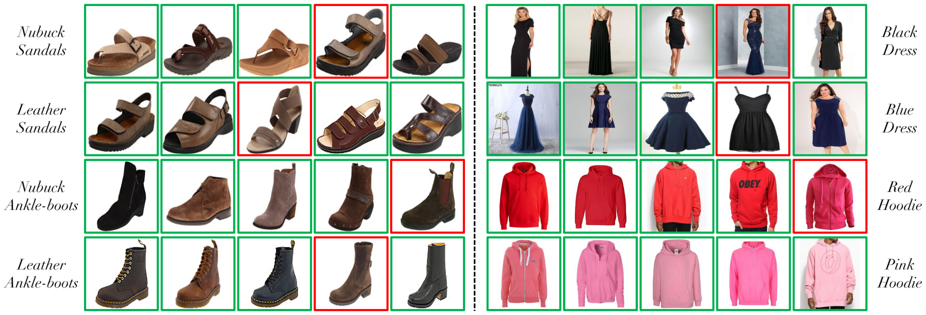

Fig. 5. Qualitative results of retrieving *nubuck sandals*, *leather sandals*, *nubuck ankle-boots*, *leather ankle-boots* in UT-Zappos50K and *black dress*, *blue dress*, *red hoodie*, *pink hoodie* in Clothing16K.

number of correct samples in the top-5, indicating that our method can solve the combinatorial generalization problem.

5 Conclusions

In reality, there are many situations where data distribution is different during training and testing. Inspired by the idea of exploring domain invariance in the DG task, we propose the representation invariant mechanism and gradient invariant mechanism to find essential features of attributes and objects, and finally learn attribute and object classifiers that can be generalized to any new composition. The limitation of our method is that it can be challenging to decouple attributes or objects when they can only form one composition in the training set. At this point, the model is more likely to overfit to the seen pairs. In the future, we will delve into studying the core features of such concepts. Besides, we will also explore the application of generalization ideas to multiple sub-concept composition scenarios and even other avenues of research.

Acknowledgement

This work was supported in part by National Natural Science Foundation of China (NSFC) No. 61922015, 62106022, U19B2036, and in part by Beijing Natural Science Foundation Project No. Z200002.

References

1. Ahuja, K., Caballero, E., Zhang, D., Gagnon-Audet, J.C., Bengio, Y., Mitliagkas, I., Rish, I.: Invariance principle meets information bottleneck for out-of-distribution generalization. NeurIPS (2021)
2. Arjovsky, M., Bottou, L., Gulrajani, I., Lopez-Paz, D.: Invariant risk minimization. arXiv preprint arXiv:1907.02893 (2019)
3. Atzmon, Y., Kreuk, F., Shalit, U., Chechik, G.: A causal view of compositional zero-shot recognition. NeurIPS (2021)
4. Bengio, Y., Deleu, T., Rahaman, N., Ke, R., Lachapelle, S., Bilaniuk, O., Goyal, A., Pal, C.: A meta-transfer objective for learning to disentangle causal mechanisms. arXiv preprint arXiv:1901.10912 (2019)
5. Blanchard, G., Lee, G., Scott, C.: Generalizing from several related classification tasks to a new unlabeled sample. NeurIPS (2011)
6. Carlucci, F.M., D'Innocente, A., Bucci, S., Caputo, B., Tommasi, T.: Domain generalization by solving jigsaw puzzles. In: CVPR (2020)
7. Chao, W.L., Changpinyo, S., Gong, B., Sha, F.: An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In: ECCV. pp. 52–68 (2016)
8. Chen, C.Y., Grauman, K.: Inferring analogous attributes. In: CVPR. pp. 200–207 (2014)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR. pp. 248–255 (2009)
10. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., Brendel, W.: Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231 (2018)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
12. Huang, Z., Wang, H., Xing, E.P., Huang, D.: Self-challenging improves cross-domain generalization. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) ECCV. pp. 124–140 (2020)
13. Isola, P., Lim, J.J., Adelson, E.H.: Discovering states and transformations in image collections. In: CVPR. pp. 1383–1391 (2015)
14. Khezeli, K., Blaas, A., Soboczenski, F., Chia, N., Kalantari, J.: On invariance penalties for risk minimization. arXiv preprint arXiv:2106.09777 (2021)
15. Kim, D., Yoo, Y., Park, S., Kim, J., Lee, J.: Selfreg: Self-supervised contrastive regularization for domain generalization. In: ICCV. pp. 9619–9628 (2021)
16. Koyama, M., Yamaguchi, S.: Out-of-distribution generalization with maximal invariant predictor. ICLR (2021)
17. Krueger, D., Caballero, E., Jacobsen, J.H., Zhang, A., Binas, J., Zhang, D., Le Priol, R., Courville, A.: Out-of-distribution generalization via risk extrapolation (rex). In: ICML. pp. 5815–5826 (2021)
18. Li, H., Pan, S.J., Wang, S., Kot, A.C.: Domain generalization with adversarial feature learning. In: CVPR (2018)
19. Li, Y.L., Xu, Y., Mao, X., Lu, C.: Symmetry and group in attribute-object compositions. In: CVPR (2020)
20. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: ECCV. pp. 852–869 (2016)
21. Mancini, M., Naeem, M.F., Xian, Y., Akata, Z.: Open world compositional zero-shot learning. In: CVPR (2021)

22. Misra, I., Gupta, A., Hebert, M.: From red wine to red tomato: Composition with context. In: CVPR. pp. 1160–1169 (2017)
23. Muandet, K., Balduzzi, D., Schölkopf, B.: Domain generalization via invariant feature representation. In: ICML. pp. 10–18 (2013)
24. Nagarajan, T., Grauman, K.: Attributes as operators: Factorizing unseen attribute-object compositions. In: ECCV (2018)
25. Palatucci, M., Pomerleau, D., Hinton, G.E., Mitchell, T.M.: Zero-shot learning with semantic output codes. NeurIPS (2009)
26. Parascandolo, G., Neitz, A., Orvieto, A., Gresele, L., Schölkopf, B.: Learning explanations that are hard to vary. arXiv preprint arXiv:2009.00329 (2020)
27. Peng, X., Huang, Z., Sun, X., Saenko, K.: Domain agnostic learning with disentangled representations. In: ICML. pp. 5102–5112 (2019)
28. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP. pp. 1532–1543 (2014)
29. Purushwalkam, S., Nickel, M., Gupta, A., Ranzato, M.: Task-driven modular networks for zero-shot compositional learning. In: ICCV (2019)
30. Qiao, F., Zhao, L., Peng, X.: Learning to learn single domain generalization. In: CVPR (2020)
31. Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N.D.: Dataset shift in machine learning. Mit Press (2008)
32. Rothenhäusler, D., Meinshausen, N., Bühlmann, P., Peters, J.: Anchor regression: Heterogeneous data meet causality. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **83**(2), 215–246 (2021)
33. Shahtalebi, S., Gagnon-Audet, J.C., Laleh, T., Faramarzi, M., Ahuja, K., Rish, I.: Sand-mask: An enhanced gradient masking strategy for the discovery of invariances in domain generalization. arXiv preprint arXiv:2106.02266 (2021)
34. Shankar, S., Piratla, V., Chakrabarti, S., Chaudhuri, S., Jyothi, P., Sarawagi, S.: Generalizing across domains via cross-gradient training. arXiv preprint arXiv:1804.10745 (2018)
35. Shi, Y., Seely, J., Torr, P.H., Siddharth, N., Hannun, A., Usunier, N., Synnaeve, G.: Gradient matching for domain generalization. arXiv preprint arXiv:2104.09937 (2021)
36. Sun, B., Saenko, K.: Deep coral: Correlation alignment for deep domain adaptation. In: ECCV. pp. 443–450 (2016)
37. Tolstikhin, I., Bousquet, O., Gelly, S., Schoelkopf, B.: Wasserstein auto-encoders. arXiv preprint arXiv:1711.01558 (2017)
38. Wang, J., Lan, C., Liu, C., Ouyang, Y., Zeng, W., Qin, T.: Generalizing to unseen domains: A survey on domain generalization. arXiv preprint arXiv:2103.03097 (2021)
39. Wei, K., Yang, M., Wang, H., Deng, C., Liu, X.: Adversarial fine-grained composition learning for unseen attribute-object recognition. In: ICCV. pp. 3741–3749 (2019)
40. Xian, Y., Lorenz, T., Schiele, B., Akata, Z.: Feature generating networks for zero-shot learning. In: CVPR. pp. 5542–5551 (2018)
41. Xian, Y., Schiele, B., Akata, Z.: Zero-shot learning-the good, the bad and the ugly. In: CVPR. pp. 4582–4591 (2017)
42. Yu, A., Grauman, K.: Fine-grained visual comparisons with local learning. In: CVPR. pp. 192–199 (2014)
43. Zhang, H., Zhang, Y.F., Liu, W., Weller, A., Schölkopf, B., Xing, E.P.: Towards principled disentanglement for domain generalization. arXiv preprint arXiv:2111.13839 (2021)

44. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
45. Zhou, K., Liu, Z., Qiao, Y., Xiang, T., Change Loy, C.: Domain generalization: A survey. arXiv preprint arXiv:2103.02503 (2021)