Adversarial Fine-Grained Composition Learning for Unseen Attribute-Object Recognition

Kun Wei[1], Muli Yang[1], Hao Wang[1], Cheng Deng[1,2]*, and Xianglong Liu[3]*
[1]School of Electronic Engineering, Xidian University, Xi'an 710071, China
[2]Tencent AI Lab, Shenzhen 518057, China
[3]State Key Lab of Software Development Environment, Beihang University, Beijing 100191, China
{weikunsk, muliyang.m, haowang.xidian, chdeng.xd}@gmail.com, xlliu@nlsde.buaa.edu.cn

Abstract

Recognizing unseen attribute-object pairs never appearing in the training data is a challenging task, since an object often refers to a specific entity while an attribute is an abstract semantic description. Besides, attributes are highly correlated to objects, i.e., an attribute tends to describe different visual features of various objects. Existing methods mainly employ two classifiers to recognize attribute and object separately, or simply simulate the composition of attribute and object, which ignore the inherent discrepancy and correlation between them. In this paper, we propose a novel adversarial fine-grained composition learning model for unseen attribute-object pair recognition. Considering their inherent discrepancy, we leverage multi-scale feature integration to capture discriminative fine-grained features from a given image. Besides, we devise a quintuplet loss to depict more accurate correlations between attributes and objects. Adversarial learning is employed to model the discrepancy and correlations among attributes and objects. Extensive experiments on two challenging benchmarks indicate that our method consistently outperforms state-of-the-art competitors by a large margin.

1. Introduction

Understanding visual concepts has always been a holy grail of computer vision. Different from supervised learning, zero-shot learning deals with the situation wherein not all samples are assigned with labels, and thus an in-depth interpretation is required to recognize the samples never appearing during training. In this paper, we consider a zero-shot recognition scenario where each sample is respectively composed of an attribute and an object (namely an adjective plus a noun). As shown in Figure 1, we train with two groups of samples *young tiger* and *old car*, and expect to successfully recognize an unseen sample of *old tiger*. It is a challenging task due to: 1) attributes and objects are intrinsically different since objects are physical entities while attributes belong to semantic descriptions, and tend to present different visual content. Therefore, it is difficult to capture consistent features for attributes and objects simultaneously and explicitly; 2) attributes are highly correlated to objects and have larger visual diversity compared with objects. For example, the attribute "old" in "old tiger" and "old car" has totally distinct visual presentations. Such imbalance between attribute and object often results in more unsatisfactory results for recognizing unseen attribute-object pairs.

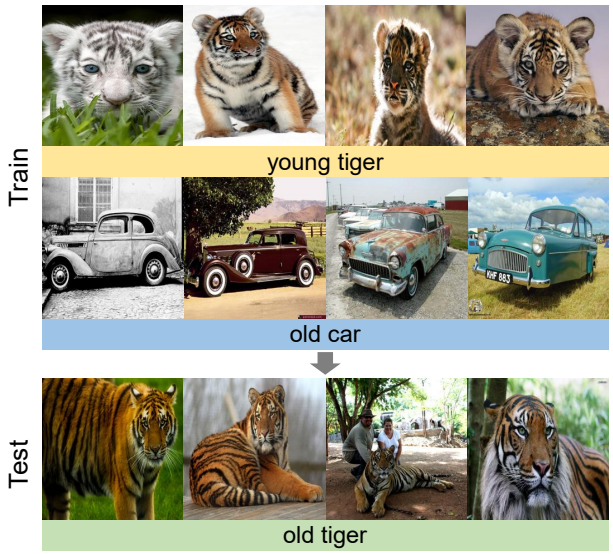


Figure 1. Overview of unseen attribute-object recognition. For example, it is expected to learn the concepts of "old" and "tiger" from the training data, and predict "old tiger" in the test data.

Traditional methods like [18] regard such problem as general recognition task by separately training classifiers for objects and attributes. These methods perform to learn attribute and object features respectively, but neglect the

*Corresponding author

inherent discrepancy and correlation between them. Intuitively, it is problematic to represent an attribute with a fixed feature due to its large visual diversity. On account of this, it is advisable to approach the intrinsic discrepancy and correlation between attributes and objects, and treat them with a unified view. Besides, other methods [19, 20] attempt to model the diverse compositions of attributes and objects, and project the compositions as well as the image visual features to a common embedding space with regularization, such as triplet loss [11, 27, 6]. However, the triplet loss, only regarding the negative sample as the one whose both attribute and object are different from the anchor, is unable to capture the complicated relationships between attribute-object pairs. In light of our observation in the experiments, these methods are easily confused by similar images (partially different images, *e.g*., "young tiger" and "old tiger") when predicting an unseen image. Therefore, it is critical to explore more fine-grained attribute-object relationships for describing the essential and subtle difference of the similar images.

In this paper, we propose an adversarial fine-grained composition learning model for recognizing unseen attribute-object pairs, aiming at establishing complete attribute-object relationships. First, we design a *quintuplet loss* to regularize the relationships among images and attribute-object pairs in a common embedding space. Unlike the triplet loss, we define the samples that are only partially different from the anchor as *semi-negative samples*. Together with the anchor, positive, semi-negative and negative samples we can construct a quintuplet for a more elaborate description of the attribute-object relationship in the common embedding space. Second, we formulate our model with GAN [9] to compose the positive and (semi-)negative samples. Benefiting from the adversarial learning, we can garner the most discriminative attribute and object features, such that the attribute-object relationships can be well-preserved and enhanced for the consequent recognition task. Third, because of the different visual presentations of attributes and objects, we find that attributes often come with details while objects usually focus on the whole concepts. Hence, we adopt multi-scale feature integration to attain more discriminative representations of attribute and object features. We evaluate our proposed method on two challenging benchmarks, *i.e*., MIT-State [12] and UT-Zappos [33]. The comparison with five state-of-the-art methods demonstrates that our method consistently achieves the best results by a large margin. Furthermore, ablation study shows that each of the adopted techniques *i.e*. fine-grained multi-scale feature integration, quintuplet loss design, and adversarial learning contributes to boosting the performance of our method.

To sum up, the contributions of this work are threefold:

- We refer to the unseen attribute-object pair recognition as a fine-grained classification task, and introduce the multi-scale feature integration to capture discriminative fine-grained features.
- A novel quintuplet loss is devised to regularize the common embedding space for an in-depth interpretation of the complicated relationships among attribute-object pairs.
- We leverage adversarial learning to build the attribute-object relationships. Different from the existing methods, we do not simply generate training samples, but flexibly compose positive and (semi-)negative pairs with an adversarial manner.

2. Related Work

Zero-Shot Learning (ZSL) is a subproblem of transfer learning, whose goal is extending supervised learning to the setting where not enough labels are available for all classes. ZSL can be extended to a more general problem, *i.e*., Generalized ZSL (GZSL), where the model is tested with both seen and unseen labels while the seen labels are excluded in ZSL. In such a setting, it is expected that we can utilize auxiliary information, *e.g*., attributes of the seen samples, to learn the composition of these attributes for the unseen ones. Recently, much progress has been made in addressing this issue with different manners, which can be typically divided into two categories: *embedding-based* methods [1, 22, 2, 25, 15, 30, 28] aiming to build a space to bridge the images and their corresponding semantic features, and *generative-based* methods [3, 5, 14] that incorporate a generative module to synthesize features of unseen categories.

Generative Adversarial Network (GAN) [9] has been involved in copious amounts of computer vision and machine learning tasks [34, 32, 31] due to its promising performance. In general, GAN involves two components, *i.e*., a generator and a discriminator. The generator learns to model the distribution of training samples and generate fake samples imitating the training ones, while the discriminator tries to distinguish the generated fake samples from the real ones. By implicitly defining the loss function with the discriminator trained adversarially along with the generator, GAN-based methods are more flexible to capture the semantic relationships between images and the corresponding class labels in zero-shot learning. Generally, most of the GAN-based ZSL methods [3, 5, 26, 8] inherit the inspiration of cGAN [17] which extends GAN from unsupervised learning into the semi-supervised setting by inputting the conditional variables along with the noise vectors. Our proposed method is also formulated with a GAN structure, but different from the existing methods, we use GAN to compose and enhance the diverse attribute-object pair relation-

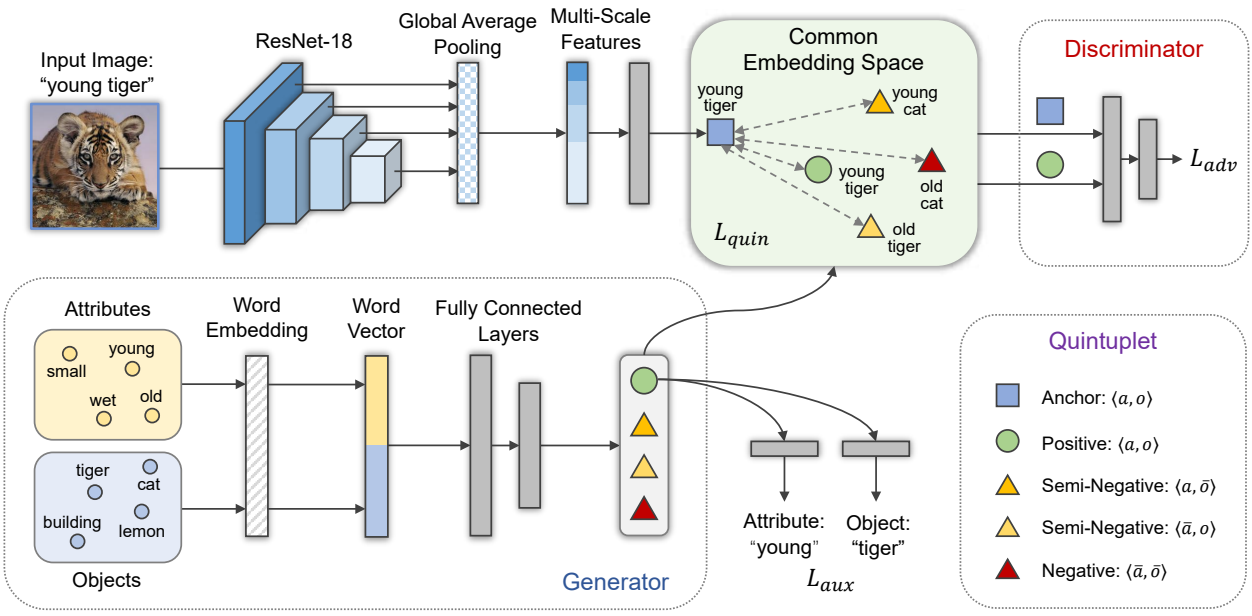


Figure 2. The framework of our proposed method, which consists of a pre-trained feature extractor, a generator and a discriminator. Given an image, the extractor captures its multi-scale features, which are then projected to a common embedding space as the anchor. Meanwhile, the generator composes four samples with concatenated attribute-object word vectors, and the positive sample is sent to a classifier with the auxiliary classification loss $\mathcal{L}_{\text{aux}}$. Then we construct a quintuplet in the common embedding space, which is regularized with the quintuplet loss $\mathcal{L}_{\text{quin}}$. The discriminator takes as input the anchor and the positive sample, and determines which input comes from the generator with the adversarial loss $\mathcal{L}_{\text{adv}}$.

ships in the common embedding space rather than simply synthesize training samples.

Unseen Attribute-Object Pair Recognition. This paper focuses on a special case of zero-shot learning scenario. Specifically, we study the situation where samples are respectively composed of an attribute (an adjective) and an object (a noun). In this setting, it is expected that we can recognize unseen attribute-object pairs with the model only trained on seen samples. To address this challenging problem, conventional methods [4, 18] utilize one or more classifiers to compose unseen attribute-object pairs with the primitive seen ones. A recent study [19] proposes to model different attributes as operators, and attribute-object pairs as objects transformed by the operators. More recently, Nan *et al*. [20] propose to find an intrinsic attribute-object representation with an encoder-decoder mechanism. In this paper, we explore the attribute-object relationships in the common embedding space, and construct the learned relationships under a GAN-based framework.

3. Methodology

The goal of this paper is to recognize the attribute-object pair of an unseen image without additional information. For instance, training with the images of "young tiger" and "old car", the expected result of the task is to correctly predict a given unseen image as "old tiger". The challenges include:

1) For zero-shot learning, the model knows little about the test attribute-object pairs that never appear in training;

2) As discussed before, this task to some extent refers to a fine-grained recognition problem. Thus, how to capture the discriminative fine-grained features is crucial;

3) The object often plays a more dominant role than the attribute, which can invalidate the recognition task.

To address these challenges, we design a quintuplet loss to regularize the composition of attribute-object pairs. We adopt GAN to adversarially compose and preserve discriminative attribute-object features. At last, fine-grained features are captured by using multi-scale feature integration.

In the following subsections, we will describe the quintuplet loss, the adversarial learning framework, the multi-scale feature integration, the overall objective function, and the training and inference procedure.

3.1. The Quintuplet Loss

Given an image $\mathbf{I}_{a,o}$, our goal is to predict its corresponding attribute-object pair label $\langle a, o \rangle$. For simplicity, we use "$^-$" to denote a negative label, *e.g.*, $\langle a, \bar{o} \rangle$ indicates a pair with the same attribute and the different object compared to $\langle a, o \rangle$. As shown in Figure 2, an image $\mathbf{I}_{a,o}$ is fed into a pre-trained feature extractor and its visual feature vector is extracted. Then the visual feature vector is projected into a common embedding space as an anchor of a triplet and de-

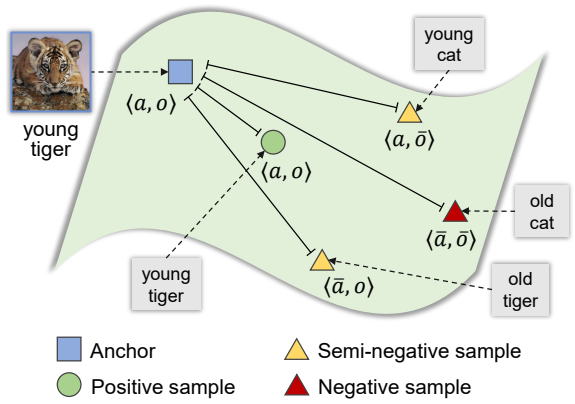


Figure 3. Illustration of the proposed quintuplet loss.

fined as $\mathbf{x}_{a,o}$, which is widely adopted in traditional methods [19, 20, 29]. Apart from the anchor $\mathbf{x}_{a,o}$, the triplet contains two other composed samples from the generator (see details in Subsection 3.2), *i.e.*, a positive sample $\hat{\mathbf{x}}_{a,o}$ and a negative one $\hat{\mathbf{x}}_{\bar{a},\bar{o}}$. The standard triplet loss ensures the anchor $\mathbf{x}_{a,o}$ to be close to the positive sample $\hat{\mathbf{x}}_{a,o}$ and far from the negative one $\hat{\mathbf{x}}_{\bar{a},\bar{o}}$, which is formulated as:

$$\mathcal{L}_{\text{triplet}}(\mathbf{x}_{a,o}, \hat{\mathbf{x}}_{a,o}, \hat{\mathbf{x}}_{\bar{a},\bar{o}}) = \max\left(0, d(\mathbf{x}_{a,o}, \hat{\mathbf{x}}_{a,o}) - d(\mathbf{x}_{a,o}, \hat{\mathbf{x}}_{\bar{a},\bar{o}}) + m\right), \quad (1)$$

where $d(\cdot,\cdot)$ denotes Euclidean distance, and m is the margin value (set to 0.5 in all involved experiments).

The triplet loss only considers $\hat{\mathbf{x}}_{\bar{a},\bar{o}}$ as the negative sample of the anchor $\mathbf{x}_{a,o}$. However, for the task of recognizing unseen attribute-object pairs, this definition for negative samples is actually insufficient. Thus, we design a quintuplet loss to regularize the common embedding space. As illustrated in Figure 3, we think that $\hat{\mathbf{x}}_{\bar{a},o}$ and $\hat{\mathbf{x}}_{a,\bar{o}}$ (*e.g.*, "young cat" and "old tiger") should lie closer to the anchor $\mathbf{x}_{a,o}$ (*e.g.*, "young tiger") than the negative sample $\hat{\mathbf{x}}_{\bar{a},\bar{o}}$ (*e.g.*, "old cat") does in the common embedding space. We also observe that large numbers of classification errors occur to the samples that are predicted as $\langle \bar{a}, o\rangle$ or $\langle a, \bar{o}\rangle$ rather than the ground truth $\langle a, o\rangle$. Therefore, we regard $\hat{\mathbf{x}}_{\bar{a},o}$ and $\hat{\mathbf{x}}_{a,\bar{o}}$ as "semi-negative samples". Together with the anchor $\mathbf{x}_{a,o}$, the positive sample $\hat{\mathbf{x}}_{a,o}$ and the negative sample $\hat{\mathbf{x}}_{\bar{a},\bar{o}}$, we can construct a quintuplet for a better depiction of the attribute-object relationships in the common embedding space. The quintuplet loss is formulated as the summation of three triplet losses:

$$\begin{aligned}\mathcal{L}_{\text{quin}}(\mathbf{x}_{a,o}, \hat{\mathbf{x}}_{a,o}, \hat{\mathbf{x}}_{\bar{a},\bar{o}}, \hat{\mathbf{x}}_{a,\bar{o}}, \hat{\mathbf{x}}_{\bar{a},o}) = \\ \lambda_1 \mathcal{L}_{\text{triplet}}\left(\mathbf{x}_{a,o}, \hat{\mathbf{x}}_{a,o}, \hat{\mathbf{x}}_{\bar{a},\bar{o}}\right) \\ + \lambda_2 \mathcal{L}_{\text{triplet}}\left(\mathbf{x}_{a,o}, \hat{\mathbf{x}}_{a,o}, \hat{\mathbf{x}}_{a,\bar{o}}\right) \\ + \lambda_3 \mathcal{L}_{\text{triplet}}\left(\mathbf{x}_{a,o}, \hat{\mathbf{x}}_{a,o}, \hat{\mathbf{x}}_{\bar{a},o}\right), \quad (2)\end{aligned}$$

where λ_1, λ_2 and λ_3 are trade-off parameters, which are respectively set to 1, 0.5, and 0.5 in all involved experiments.

3.2. Adversarial learning for Composition Pairs

We construct a GAN to model the composition of attribute-object pairs and enhance the attribute-object relationships with adversarial learning. GAN consists of a generator G and a discriminator D, where G is to compose the attribute-object pairs and D is to distinguish whether a pair is composed by the generator G.

In particular, the generator G takes as input the attribute word vector $\mathbf{w}_a$ and the object word vector $\mathbf{w}_o$ corresponding to the anchor $\mathbf{x}_{a,o}$. Then, the two word vectors are concatenated and projected to the common embedding space as a composed attribute-object pair vector, which is defined as a positive sample $\hat{\mathbf{x}}_{a,o}$. Accordingly, with different input word vectors ($\mathbf{w}_a$, $\mathbf{w}_{\bar{a}}$, $\mathbf{w}_o$, and $\mathbf{w}_{\bar{o}}$), the generator composes $\hat{\mathbf{x}}_{a,\bar{o}}$, $\hat{\mathbf{x}}_{\bar{a},o}$ and $\hat{\mathbf{x}}_{\bar{a},\bar{o}}$ as the semi-negative and negative samples. The discriminator D takes as input the anchor $\mathbf{x}_{a,o}$ and the composed positive sample $\hat{\mathbf{x}}_{a,o}$, and determines which input is produced by the generator G. The discriminator D is designed as a multi-layer perceptron, which promotes the generator G to compose discriminative attribute-object features with the overall adversarial loss:

$$\begin{aligned}\mathcal{L}_{\text{adv}} = \mathbb{E}_{\mathbf{x}_{a,o}}\left(\log D\left(\mathbf{x}_{a,o}\right)\right) \\ + \mathbb{E}_{\mathbf{w}_a,\mathbf{w}_o}\left(\log\left(1 - D\big(G\left(\mathbf{w}_a, \mathbf{w}_o\right)\big)\right)\right), \quad (3)\end{aligned}$$

where $G\left(\mathbf{w}_a, \mathbf{w}_o\right) = \hat{\mathbf{x}}_{a,o}$, and G tries to minimize $\mathcal{L}_{\text{adv}}$ while D tries to maximize it.

3.3. Multi-Scale Feature Integration

As discussed in Section 1, recognizing the unseen attribute-object pairs requires fine-grained discriminative attribute features. In fact, the commonly-used feature extractor (ResNet-18 [10]) is pre-trained on the ImageNet dataset [23], which is collected for the object recognition task. As a result, the extracted visual feature vector contains considerably more object features than the attribute ones, such that the visual features of images $\mathbf{I}_{a,o}$ and $\mathbf{I}_{\bar{a},o}$ (*e.g.*, "young tiger" and "old tiger") can be extremely similar. To address this problem, we introduce the multi-scale feature integration. Without fine-tuning the pre-trained feature extractor, the features are fused from layers of different depths as shown in Figure 2. Compared with the features only extracted from the last layer, the features from lower ones contain more fine-grained information, which is beneficial to acquiring more discriminative visual features. Through global average pooling and concatenating, we can derive the final visual feature vector. The effectiveness of the multi-scale feature integration will be proven in the experiment part in Section 4.

3.4. Overall Objective Function

In our method, the composition of attribute-object pairs is flexible. There is always a possibility that either an attribute or an object plays a prominent role in the composition process, which may bring about an imbalance between attributes and objects. Such imbalance often leads to the high classification accuracy for the major one and low for the other one, which can cause the overall attribute-object prediction inaccurate. Therefore, we introduce an auxiliary classification loss that guides the composition process, which is formulated as:

$$\mathcal{L}_{\text{aux}} = h_a(\hat{\mathbf{x}}_{a,o}, a) + h_o(\hat{\mathbf{x}}_{a,o}, o)\,, \tag{4}$$

where $h_a(\cdot)$ and $h_o(\cdot)$ are both a fully-connected layer with the cross-entropy loss trained to classify attributes and objects respectively. The feature of attribute and object can be reserved in the composition with the supervision of the auxiliary classification loss.

Finally, the objectives of G and D are written as:

$$\mathcal{L}_D = -\,\mathcal{L}_{\text{adv}}\,, \tag{5}$$

$$\mathcal{L}_G = \mathcal{L}_{\text{adv}} + \lambda_{\text{quin}}\mathcal{L}_{\text{quin}} + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}}\,, \tag{6}$$

where λ_{quin} and λ_{aux} are trade-off parameters, which are respectively set to 1 and 1000 in all involved experiments.

3.5. Training and Inference

During training, we project each image $\mathbf{I}_{a,o}$ into a common embedding space as the anchor $\mathbf{x}_{a,o}$. The attribute and object word vectors $\mathbf{w}_a$, $\mathbf{w}_{\bar{a}}$, $\mathbf{w}_o$, and $\mathbf{w}_{\bar{o}}$ are concatenated and projected into the common embedding space as four samples $\hat{\mathbf{x}}_{a,o}$, $\hat{\mathbf{x}}_{a,\bar{o}}$, $\hat{\mathbf{x}}_{\bar{a},o}$, and $\hat{\mathbf{x}}_{\bar{a},\bar{o}}$ by the generator G. We note that $\mathbf{w}_{\bar{a}}$ and $\mathbf{w}_{\bar{o}}$ are randomly selected to be different from $\mathbf{w}_a$ and $\mathbf{w}_o$. The quintuplet loss pulls $\hat{\mathbf{x}}_{a,o}$ close to $\mathbf{x}_{a,o}$, and $\hat{\mathbf{x}}_{a,\bar{o}}$, $\hat{\mathbf{x}}_{\bar{a},o}$, $\hat{\mathbf{x}}_{\bar{a},\bar{o}}$ away from $\mathbf{x}_{a,o}$. The losses guarantee the discrimination of generated composition for promising classification results.

For inference, given an unseen image $\mathbf{I}$ with the vector $\mathbf{x}$ in common embedding space, we generate the compositions of candidate pairs from all available word vectors. The distances between $\mathbf{x}$ and each candidate pair $\hat{\mathbf{x}}$ are computed and sorted. Then, the candidate pair $\hat{\mathbf{x}}_{a,o}$ corresponding to the shortest distance is regarded as the prediction, *i.e.*, the unseen image $\mathbf{I}$ is predicated as $\mathbf{I}_{a,o}$.

4. Experiments

In this section, all involved datasets, evaluation metrics and baselines are introduced in detail. Then, we will present the implementation details as well as the experimental results of our method and several state-of-the-art competitors. Finally, two ablation studies will prove the effectiveness of our proposed approach.

4.1. Datasets

We evaluate our method on two popular datasets.

MIT-States [12] has a wide range of objects and attributes. It contains 245 object classes, 115 attribute classes, with 53,753 images in total. Each image is annotated with an attribute-object pair such as "young tiger". Since not all pairs make sense in the real world, it contains 1962 attribute-object pairs rather than 28,175 pairs. We use the compositional split [18], *i.e.*, 1262 pairs (34,562 images) for training and 700 pairs (19,191 images) for testing. The training pairs and testing pairs are non-overlapping.

UT-Zappos [33] contains 50,025 images of shoes with attribute labels, which has 16 attribute classes and 12 object classes. Following the same setting in [19], we use 83 attribute-object pairs (24,898 images) for training and 33 pairs for testing (4228 images).

4.2. Evaluation

We evaluate the methods by Top-1 accuracy on recognizing unseen attribute-object pairs. The accuracy is reported via three metrics:

Closed: testing pair candidates are restricted to the unseen pairs. At the test stage, we measure the embedding distances between a given image and only the unseen pairs, then predict the image as the nearest composed pair. The Closed metric reduces the number of testing candidates and usually achieves better accuracy, but is not practical for real-world applications.

Open: testing pair candidates are open for all seen and unseen pairs. During testing, we consider both seen and unseen composed pairs as candidates for recognition, which is more practical and challenging. The embedding distances are measured between a given image and the pair candidates, and then the image is predicted as the nearest composed pair.

H-Mean: Harmonic Mean measures the overall performance of both Closed and Open metrics, defined as:

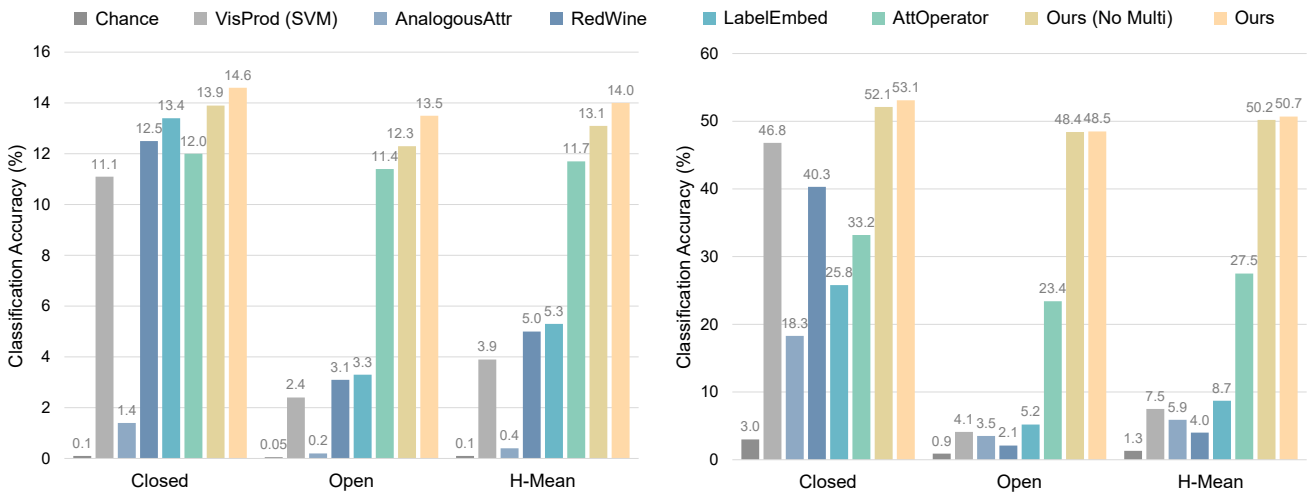
$$A_{\text{H}} = 2 \times \frac{A_{\text{C}} \times A_{\text{O}}}{A_{\text{C}} + A_{\text{O}}}\,, \tag{7}$$

where A_{H}, A_{C}, and A_{O} respectively denote the accuracy with H-Mean, Closed, and Open metrics. As a broadly used evaluation metric [14, 26, 5, 24], Harmonic Mean balances the performance between the Closed and Open metrics.

4.3. Compared Baselines

Our method is compared with the following baselines:

VisProd [16] trains two classifiers to predict the attribute and object separately. The Linear SVM is employed as the classifiers, and the overall accuracy is calculated as the product of the separate accuracy for attributes and objects.



(a) MIT-States. (b) UT-Zappos.

Figure 4. Classification accuracy of unseen pair recognition with the three evaluation metrics on the two datasets. We note that "Chance" indicates random prediction.

AnalogousAttr [4] trains a Linear SVM classifier on seen attribute-object pairs, and predicts unseen pairs with the trained classifier.

RedWine [18] uses pre-trained classifier weights (Linear SVM) to compose word vector representations, and trains a neural network to recognize unseen attribute-object pairs.

LabelEmbed [7] uses pre-trained GloVe [21] word embeddings to compose word vector representations, which is the difference compared with RedWine.

AttOperator [19] regards attributes as operators, and simulates attribute-object pair compositions as attribute-conditioned transformations. The training and testing pairs are also non-overlapping.

4.4. Implementation

For each image, we extract a 960-dimensional multi-scale visual feature vector using ResNet-18 [10] pre-trained on the ImageNet dataset [23]. For each attribute-object pair, we extract a 960-dimensional linguistic feature vector for both attribute and object with word embeddings. Our model is implemented with PyTorch[1] and optimized by ADAM optimizer [13] on an NVIDIA GTX 1080Ti GPU. The learning rate and batch size are respectively set to 0.0001 and 512. For the MIT-States dataset, the training time is approximately 5h for 1000 epochs. For the UT-Zappos dataset, it takes around 2h for 1000 epochs in training.

4.5. Results and Analysis

Figure 4 shows the results of our method compared with the baselines. Our method consistently outperforms all the baselines with all the metrics by a large margin. On the MIT-States dataset, compared with state of the art, our method increases the classification accuracy by 1.2% (Closed), 2.1% (Open), and 2.3% (H-Mean). On the UT-Zappos dataset, the accuracy of our method increases by 6.4% (Closed), 25.1% (Open), and 23.2% (H-Mean). The experimental result sufficiently proves the superiority of our proposed method. Since the baselines directly use visual features extracted from the last layer of ResNet-18, for fair comparisons, we also present the result of our method without multi-scale feature integration (denoted as "No Multi"). Our classification accuracy is slightly worse than the final ones but still better than the baselines.

Compared with UT-Zappos, MIT-States has a much larger number of attributes, objects, unseen pairs, more complex backgrounds, and fewer training samples for each attribute-object pair, and thus is more difficult to learn robust composition for unseen pairs. Therefore, MIT-States benefits relatively less than UT-Zappos.

With the Closed metric, it is relatively easy to produce satisfactory results by artificially decreasing the number of pair label candidates. As shown in Figure 4, except AttOperator and ours, all other methods perform considerably worse with the challenging Open metric than the Closed one, which indicates over-fitting on the datasets.

Figure 5 shows some qualitative results on the two datasets. On the MIT-States dataset, our method is able to recognize some samples like "Ancient Building", "Cluttered Cabinet" and "Huge Dog", which present obvious attributes and objects. For "Broken Laptop", the attribute "Broken" is relatively apparent whereas the object "Laptop" can vary to a large extent, which causes the error predictions as $\langle a, \bar{o} \rangle$. As for "Cored Apple", the appearance of

[1] https://pytorch.org/

(a) Images in MIT-States.

(b) Images in UT-Zappos.

Figure 5. Qualitative results of our method on the two datasets. For each dataset, the left three columns (marked in green) show the samples with correct predictions $\langle a, o\rangle$, the next two columns (in orange) show the samples with predictions as $\langle a, \bar{o}\rangle$ and $\langle \bar{a}, o\rangle$ respectively, and the last column (in red) shows the samples with false predictions $\langle \bar{a}, \bar{o}\rangle$.

Table 1. Ablation study: classification accuracy (%) with three different modules. "$\mathcal{L}_{\text{quin}}$", "$\mathcal{L}_{\text{adv}}$" and "Multi" respectively indicate the quintuplet loss, adversarial learning and multi-scale feature integration.

	MIT-States			UT-Zappos		
Method	Closed	Open	H-Mean	Closed	Open	H-Mean
Base	13.4	4.5	6.7	37.4	9.4	15.0
+ $\mathcal{L}_{\text{quin}}$	12.8	11.3	12.0	49.9	47.7	48.8
+ $\mathcal{L}_{\text{adv}}$	14.0	4.4	6.7	43.4	16.0	23.3
+ Multi	13.9	5.3	7.7	47.8	11.2	18.2
+ $\mathcal{L}_{\text{quin}}$ + $\mathcal{L}_{\text{adv}}$	13.9	12.3	13.1	52.1	48.4	50.2
+ $\mathcal{L}_{\text{quin}}$ + Multi	14.1	12.9	13.5	51.0	47.8	49.3
+ $\mathcal{L}_{\text{adv}}$ + Multi	**15.0**	5.8	8.4	52.5	16.3	24.9
+ $\mathcal{L}_{\text{quin}}$ + $\mathcal{L}_{\text{adv}}$ + Multi	14.6	**13.5**	**14.0**	**53.1**	**48.5**	**50.7**

"Apple" is obvious, but "Cored" is various and visually hard to understand. As for "Diced Fruit", there can be many objects in different states mixed in the images, which is a big challenge for our method. Besides, some attributes are confusable and some objects have similar appearances, which is one of the main reasons leading the false prediction. For example, the attribute "Old" is similar to "Ancient", the object "Cat" is similar to "Tiger". On the UT-Zappos dataset, as the number of attribute-object pairs is small, the results are relatively better. We can observe that some attributes or objects with indistinctive visual cues can be recognized wrongly, such as " Sheepskin" and "Oxfords Shoes".

4.6. Ablation Study

We conduct two groups of experiments to study the effectiveness of the three modules.

The results of our base model adding different modules are presented in Table 1. The base model is constrained by the triplet loss, which is to be replaced by our quintuplet loss when adding "$\mathcal{L}_{\text{quin}}$". As shown in Table 1, both adversarial learning (denoted as "$\mathcal{L}_{\text{adv}}$") and multi-scale feature integration (denoted as "Multi") can improve the performance with all Closed, Open and H-Mean metrics on the two datasets. The improvement of adding "Multi" indicates the multi-scale feature integration is able to capture more discriminative fine-grained visual features by fusing the outputs of layers in different depth. The improvement of adding "$\mathcal{L}_{\text{adv}}$" demonstrates that adversarial training is helpful to compose attribute-object pairs and preserve discriminative features. However, these two modules cannot minimize the performance disparity between Closed and Open metrics, which indicates the model still over-fits the

Table 2. Ablation study: classification accuracy (%) with the triplet/quintuplet loss on MIT-States. "Attribute-Object", "Attribute", and "Object" respectively indicate the performances for predicting the attribut-object pairs, only attributes, and only objects.

Method	Closed			Open		
	Attribute-Object	Attribute	Object	Attribute-Object	Attribute	Object
Ours (with $\mathcal{L}_{\text{triplet}}$)	**15.0**	23.0	**25.9**	5.8	16.4	24.3
Ours (with $\mathcal{L}_{\text{quin}}$)	14.6	**23.4**	24.4	**13.5**	**22.0**	**24.9**

Table 3. Ablation study: classification accuracy (%) with the triplet/quintuplet loss on UT-Zappos. "Attribute-Object", "Attribute", and "Object" respectively indicate the performances for predicting the attribut-object pairs, only attributes, and only objects.

Method	Closed			Open		
	Attribute-Object	Attribute	Object	Attribute-Object	Attribute	Object
Ours (with $\mathcal{L}_{\text{triplet}}$)	52.2	55.5	77.3	16.3	30.4	67.2
Ours (with $\mathcal{L}_{\text{quin}}$)	**53.1**	**56.2**	**78.4**	**48.5**	**52.5**	**78.4**

Table 4. Ablation study: numbers of partly correct predictions on MIT-States. We note that there are 19,191 images in the test set.

Method	Closed		Open	
	$\langle \bar{a}, o \rangle$	$\langle a, \bar{o} \rangle$	$\langle \bar{a}, o \rangle$	$\langle a, \bar{o} \rangle$
Ours (with $\mathcal{L}_{\text{triplet}}$)	2096	1616	3536	2029
Ours (with $\mathcal{L}_{\text{quin}}$)	1875	1598	2191	1632

Table 5. Ablation study: numbers of partly correct predictions on UT-Zappos. We note that there are 4228 images in the test set.

Method	Closed		Open	
	$\langle \bar{a}, o \rangle$	$\langle a, \bar{o} \rangle$	$\langle \bar{a}, o \rangle$	$\langle a, \bar{o} \rangle$
Ours (with $\mathcal{L}_{\text{triplet}}$)	1047	128	2152	593
Ours (with $\mathcal{L}_{\text{quin}}$)	1072	130	1265	168

two datasets. As adding "$\mathcal{L}_{\text{quin}}$" to the base model, the accuracy with Closed and Open metrics is balanced, which implies the quintuplet loss can regularize the common embedding space effectively. Finally, we combine the three modules and achieve the best results with all the three evaluation metrics on the two datasets.

We also conduct another experiment to evaluate the capacity of our method to resist the situation predicting $\mathbf{I}_{a,o}$ as $\langle a, \bar{o} \rangle$ or $\langle \bar{a}, o \rangle$. In Table 2 and 3, we report the results (accuracy of classifying attribute-object pair, only the attribute, and only the object) of our method on the two datasets with the triplet/quintuplet loss. We notice that the accuracy with the Closed metric seems not influenced by replacing the triplet loss with the quintuplet one, the reason of which is that the number of the interference candidates (*i.e.*, $\langle a, \bar{o} \rangle$ or $\langle \bar{a}, o \rangle$) is very small. This situation is opposite with the Open metric where the accuracy increases by a large margin. As listed in Table 4 and 5, we count the images predicted as $\langle a, \bar{o} \rangle$ or $\langle \bar{a}, o \rangle$ rather than the ground truth $\langle a, o \rangle$ with the two losses. As replacing the triplet loss with the quintuplet one, we see a considerable decrease of partly correct predictions with the Open metric on both the two datasets. We can infer from Table 2, 3, 4, and 5 that the classification accuracy of only attribute (or only object) increases just following the pattern of which partly correct predictions decrease. Thus, we can conclude that our proposed quintuplet loss is able to resist the interference from partly correct samples.

5. Conclusion

In this paper, we have proposed a novel adversarial fine-grained composition learning model to recognize unseen attribute-object pairs. We design a quintuplet loss to regularize the common embedding space, achieving a better interpretation of the inherent and complex attribute-object relationships. The adversarial learning strategy is leveraged to model the composition of attributes and objects and preserve attribute-object relationships. We introduce the multi-scale feature integration to acquire more discriminative fine-grained features. Experiments show our method outperforms state-of-the-art competitors by a large margin on two benchmark datasets with all Closed, Open, and Harmonic Mean metrics. In the future, we plan to continue investigating the relationships between attributes and objects, and cope with the compositions involving multiple attributes and objects.

Acknowledgment

Our work was supported in part by the National Natural Science Foundation of China under Grant 61572388 and 61703327, in part by the Key R&D Program-The Key Industry Innovation Chain of Shaanxi under Grant 2017ZDCXL-GY-05-04-02, 2017ZDCXL-GY-05-02 and 2018ZDXM-GY-176, and in part by the National Key R&D Program of China under Grant 2017YFE0104100.

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE Trans. Pattern Anal. Machine Intell.*, 38(7):1425–1438, 2016.

[2] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, pages 2927–2936, 2015.

[3] Maxime Bucher, Stéphane Herbin, and Frédéric Jurie. Generating visual representations for zero-shot classification. In *ICCV*, pages 2666–2673, 2017.

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

[5] Long Chen, Hanwang Zhang, Jun Xiao, Wei Liu, and Shih-Fu Chang. Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In *CVPR*, pages 1043–1052, 2018.

[6] Cheng Deng, Zhaojia Chen, Xianglong Liu, Xinbo Gao, and Dacheng Tao. Triplet-based deep hashing network for cross-modal retrieval. *IEEE Trans. Image Process.*, 27(8):3893–3903, 2018.

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

[8] Rafael Felix, Vijay BG Kumar, Ian Reid, and Gustavo Carneiro. Multi-modal cycle-consistent generalized zero-shot learning. In *ECCV*, pages 21–37, 2018.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, pages 2672–2680, 2014.

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

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

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

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

[14] Vinay Kumar Verma, Gundeep Arora, Ashish Mishra, and Piyush Rai. Generalized zero-shot learning via synthesized examples. In *CVPR*, pages 4281–4289, 2018.

[15] Yan Li, Junge Zhang, Jianguo Zhang, and Kaiqi Huang. Discriminative learning of latent features for zero-shot recognition. In *CVPR*, pages 7463–7471, 2018.

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

[17] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

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

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

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

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

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

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

[24] Qian Wang and Ke Chen. Alternative semantic representations for zero-shot human action recognition. In *ECML-PKDD*, pages 87–102. Springer, 2017.

[25] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *CVPR*, pages 69–77, 2016.

[26] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *CVPR*, pages 5542–5551, 2018.

[27] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *CVPR*, pages 4582–4591, 2017.

[28] Xinyi Xu, Cheng Deng, and Feiping Nie. Adaptive graph weighting for multi-view dimensionality reduction. *Signal Processing*, 2019.

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

[30] Muli Yang, Cheng Deng, and Feiping Nie. Adaptive-weighting discriminative regression for multi-view classification. *Pattern Recognition*, 88:236–245, 2019.

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

[32] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, pages 2849–2857, 2017.

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

[34] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *ICCV*, 2017.