International Journal of Automation and Computing 16(5), October 2019, 563-574
DOI: 10.1007/s11633-019-1177-8

Zero-shot Fine-grained Classification by Deep Feature Learning with Semantics

Ao-Xue Li Ke-Xin Zhang Li-Wei Wang

The Key Laboratory of Machine Perception (MOE), School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China

Abstract: Fine-grained image classification, which aims to distinguish images with subtle distinctions, is a challenging task for two main reasons: lack of sufficient training data for every class and difficulty in learning discriminative features for representation. In this paper, to address the two issues, we propose a two-phase framework for recognizing images from unseen fine-grained classes, i.e., zero-shot fine-grained classification. In the first feature learning phase, we finetune deep convolutional neural networks using hierarchical semantic structure among fine-grained classes to extract discriminative deep visual features. Meanwhile, a domain adaptation structure is induced into deep convolutional neural networks to avoid domain shift from training data to test data. In the second label inference phase, a semantic directed graph is constructed over attributes of fine-grained classes. Based on this graph, we develop a label propagation algorithm to infer the labels of images in the unseen classes. Experimental results on two benchmark datasets demonstrate that our model outperforms the state-of-the-art zero-shot learning models. In addition, the features obtained by our feature learning model also yield significant gains when they are used by other zero-shot learning models, which shows the flexility of our model in zero-shot fine-grained classification.

Keywords: Fine-grained image classification, zero-shot learning, deep feature learning, domain adaptation, semantic graph.

1 Introduction

Fine-grained image classification, which aims to recognize subordinate level categories, has emerged as a popular research area in the computer vision community[1–5]. Different from general image recognition such as scene or object recognition, fine-grained image classification needs to explicitly distinguish images with subtle difference, which actually involves the classification of many subclasses of objects belonging to the same class such as birds[6–8], dogs[9] and plants[10, 11].

In general, fine-grained image classification is a challenging task due to two main issues:

1) Since recognizing images in the fine-grained classes is a fairly difficult and expert task, the annotations of images in fine-grained classes are expensive and collecting large-scale labelled data just as general image recognition (e.g., ImageNet[12]) is thus impractical. Therefore, the question of how to recognize images from fine-grained classes given the lack of sufficient training data for every class becomes a thought-provoking one in computer vision.

2) As compared with general image recognition, fine-grained classification is a more challenging task, which needs to discriminate between objects that are visually similar to each other. Therefore, we have to learn more discriminative representation for fine-grained classification than that for general image classification.

Considering the lack of training data for every class in fine-grained classification, we can adopt zero-shot learning to recognize images from unseen classes without labelled training data. However, conventional zero-shot learning algorithms mainly explore the semantic relationship among classes (using textual information) and attempt to learn a match between images and their textual descriptions[13–15]. In other words, this rarely works on zero-shot learning focus on feature learning. This is really bad for fine-grained classification, since it requires more discriminative features than general image recognition. Hence, we must focus on feature learning for zero-shot fine-grained image classification.

In this paper, we propose a two-phase framework to recognize images from unseen fine-grained classes, i.e., zero-shot fine-grained classification (ZSFC). The first phase of our model is to learn discriminative features. Most fine-grained classification models extract features from deep convolutional neural networks that are fine-tuned by images with extra annotations (e.g., bounding box of objects and part locations). However, these extra annotations of images are expensive to access. Unlike these models, our model only exploits implied hierarchical semantic structure among fine-grained classes for fine-tuning deep networks. The hierarchical semantic struc-

Research Article
Manuscript received October 10, 2018; accepted March 8, 2019; published online May 15, 2019
Recommended by Associate Editor Bin Luo

The original version of this article was revised due to a retrospective Open Access order

ture among classes is obtained based on taxonomy, which can be easily collected from Wikipedia. In our model, we generally assume that experts recognize objects in fine-grained classes based on the discriminative visual features of images and the hierarchical semantic structure among fine-grained classes is their prior knowledge. Under this assumption, we finetune deep convolutional neural networks using hierarchical semantic structure among fine-grained classes to extract discriminative deep visual features. Meanwhile, a domain adaptation subnetwork is introduced into the proposed network to avoid domain shift caused by zero-shot setting.

In the second label inference phase, a semantic directed graph is firstly constructed over attributes of fine-grained classes. Based on the semantic directed graph and also the discriminative features obtained by our feature learning model, we develop a label propagation algorithm to infer the labels of images in the unseen classes. The flowchart of the proposed framework is illustrated in Fig. 1. Note that the proposed framework can be extended to a weakly supervised setting by replacing class attributes with semantic vectors extracted by word vector extractors (e.g., Word2Vec[16]).

To evaluate the effectiveness of the proposed model, we conduct experiments on two benchmark fine-grained image datasets (Caltech UCSD Birds-200-2011[6] and Oxford Flower-102[10]). Experimental results demonstrate that the proposed model outperforms the state-of-the-art zero-shot learning models in the task of zero-shot fine-grained classification. Moreover, we further test the features extracted by our feature learning model by applying them to other zero-shot learning models and the obtained significant gains verify the effectiveness of our feature learning model.

The main contributions of this work are given as follows:

1) We have proposed a two-phase learning framework for zero-shot fine-grained classification. Unlike most previous works that focus on zero-shot learning, we pay more attention to feature learning instead.

2) We have developed a deep feature learning method for fine-grained classification, which can learn discriminative features with hierarchical semantic structure among classes and a domain adaptation structure. More notably, our feature learning method needs no extra annotations of images (e.g., part locations and bounding boxes of objects), which means that it can be readily used for different zero-shot fine-grained classification tasks.

3) We have developed a zero-shot learning method for label inference from seen classes to unseen classes, which can help to address the issue of lack of labelled training data in fine-grained image classification.

The remainder of this paper is organized as follows. Section 2 provides related works of fine-grained classification and zero-shot learning. Section 3 gives the details of the proposed model for zero-shot fine-grained classification. Experimental results are presented in Section 4. Finally, the conclusions are drawn in Section 5.

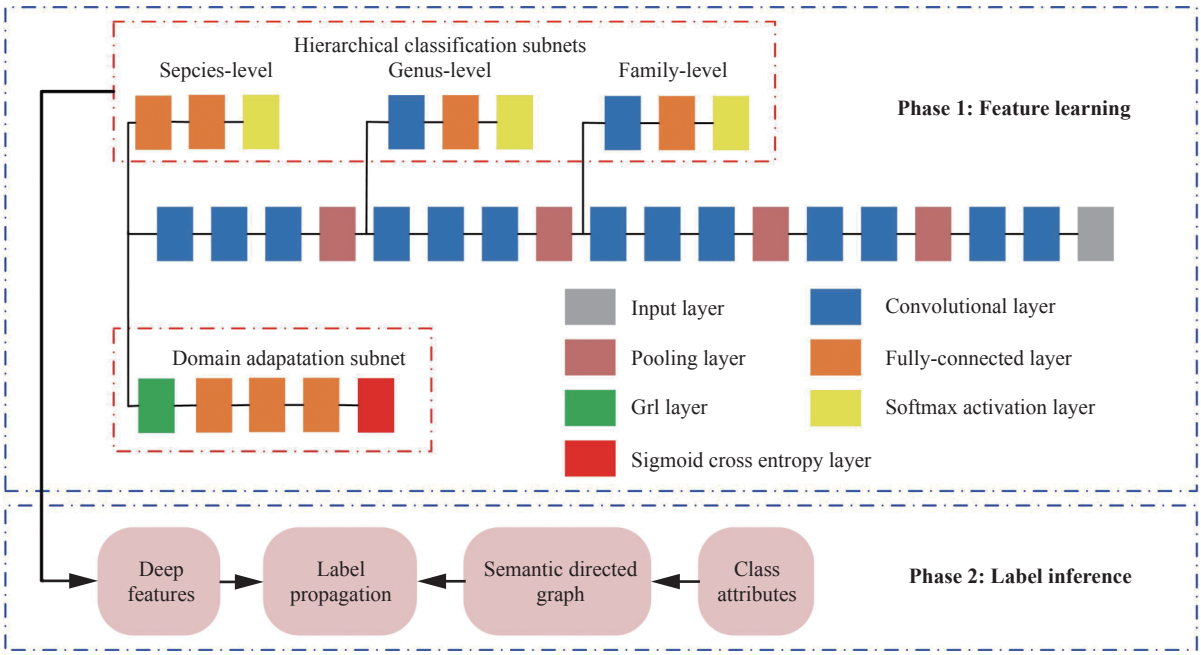


Fig. 1 Overview of the proposed framework for zero-shot fine-grained image classification. The proposed framework contains two phases: feature learning and label inference. In the first feature learning phase, hierarchical classification subnetworks and a domain adaptation structure are both integrated into VGG-16Net. In the second label inference phase, deep features from the first phase and a semantic directed graph constructed with class attributes are involved into a label propagation process to infer the labels of images in the unseen classes. Color versions of the figures in this paper are available online.

2 Related works

2.1 Fine-grained image classification

There are two strategies widely used in existing fine-grained image classification algorithms. The idea of the first strategy is distinguishing images according to the unique properties of object parts, which encourages the use of part-based algorithms that rely on localizing object parts and assigning them region-based convolutional neural network attributes. Zhang et al.[17] propose a part-based region based-convolutional neural network (R-CNN) where R-CNN is used to detect object parts and geometric relations among object parts which are used for label inference. Since R-CNN extracts too many proposals for each image, this algorithm is time-consuming. To solve this problem, Huang et al.[18] propose a part-stacked convolutional neural network (PS-CNN), where a fully-convolutional network is used to detect object parts and a part-crop layer is induced into AlexNet[19] to combine part/object features for classification. To solve the limited scale of well-annotated data, Xu et al.[20] propose an agumented part-based R-CNN to utilize the weak labeled data from the web. Unlike those models that mainly use large parts of images (i.e., proposals) for fine-grained classification, Zhang et al.[21] detect semantic parts and classify images based on features of their semantic parts. However, the aforementioned part-based algorithms need very strong annotations (i.e., locations of parts), which are very expensive to acquire.

The second strategy is to exploit more discriminative visual representations, which is inspired by the recent success of CNNs in image recognition[22]. Lin et al.[23] propose a bilinear CNN, which combines the outputs of two different feature extractors by using an outer product to model local pairwise feature interactions in a translationally invariant manner. This structure can create robust representations and achieve significant improvement compared with the state-of-the-art. Zhang et al.[24] propose a deep filter selection strategy to choose suitable deep filters for each kinds of parts. With suitable deep filters, they can detect more accurate parts and extract more discriminative features for fine-grained classification.

Note that the above models need extra annotations of images (e.g., bounding boxes of objects and locations of parts). Moreover, their training data include all fine-grained classes. When we only have training images from a subset of fine-grained classes, the domain shift problem will occur[25]. Besides, without extra object or part annotations, these models will fail. In contrast, our model needs no extra object or part annotations at both training and testing stages. Furthermore, the domain adaptation strategy is induced into our model to avoid domain shift. In this way, we can learn more discriminative features for zero-shot fine-grained classification.

2.2 Zero-shot learning

Zero-shot learning, which aims to learn to classify in the absence of labeled data, is a challenging problem[26–32]. Recently, many zero-shot learning approaches have been developed. Zhang and Saligrama[14] viewed each source or target data as a mixture of seen class proportions and postulated that the mixture patterns have to be similar if the two instances belong to the same unseen class. A semantic similarity embedding (SSE) approach for zero-shot learning is proposed to solve this problem. They also formulate zero-shot learning as a binary classification problem and develop a joint discriminative learning framework based on dictionary learning to solve it[33]. Romera-Paredes and Torr[13]. use a two linear layers network to model the relationships between features, attributes, and classes. Bucher et al.[34] addresses the task of zero-shot learning by formulating this problem as a metric learning problem, where a metric among class attributes and image visual features is learned for inferring labels of test images. A multi-cue framework facilitates a joint embedding of multiple language parts and visual information into a joint space to recognize images from unseen classes[35]. Fu et al.[15] propose to model the semantic mainfold in an embedding space using a semantic class label graph, in order to redefine the distance metric in the semantic embedding space for more effective zero-shot learning (ZSL). To avoid domain shift between the sets of seen classes and unseen classes, Kodirov et al.[25] propose a zero-shot learning method based on unsupervised domain adaptation. On the observation that textual descriptions are noisy, Qiao et al.[36] propose an $L_{2,1}$-norm based objective function to suppress the noisy signal in the text and provide a function to match the text document and visual features of images. However, the aforementioned works mainly focus on recognizing a match between images and their textual descriptions and few of them pay attention to discriminative feature learning, which is crucial for fine-grained classification.

3 Proposed model

In this section, we propose a two-phase framework for zero-shot fine-grained classification. A deep convolutional neural network integrating hierarchical semantic structure of classes and domain adaptation strategy is first developed for feature learning and a label propagation method based on semantic directed graphs is further proposed for label inference.

3.1 Feature learning

Our main idea is motivated by implied hierarchical semantic structure among fine-grained classes. For example, winter wren (species-level name), a very small North-American bird, can be called “Troglodytes” at genus level and also can be called “Troglodytidae” at family level

(See Fig. 2). We assume that experts recognize objects in fine-grained classes by using the discriminative visual features and the hierarchical semantic structure among fine-grained classes is their prior knowledge. As shown in Fig. 1, lower-level features are used (with fewer network layers) for classifying images at coarser level. In other words, to recognize images in a fine-grained level, we must exploit higher-level and fine-grained features.

To induce the hierarchical semantic structure into feature learning, we integrate hierarchical classification subnetworks (HCS) into VGG-16Net[37]. The detailed architectures of hierarchical classification subnetworks are presented in Fig. 3. In our model, each classification subnetwork is designed to classify images into the corresponding level semantic classes (i.e., family level, genus level, or species level). Concretely, we locate the classification subnetworks for family-level, genus-level, and species-level labels afterwards the third, fourth, and fifth groups of convolutional layers, respectively (also see Fig. 1). For family-level and genus-level classification subnetworks, their detailed network structure includes a convolutional layer, two fully-connected layers, and a softmax activation layer (see Fig. 3). For the sake of quick converegence, we take the classification structure of VGG-16Net as the species-level classification subnetwork (see Fig. 3), which can be initialized by ImageNet pre-trained parameters[12]. Therefore, by merging the VGG-16Net and hierarchical classification subnetworks into one network, we define the loss function for an image x as follows.

$$\begin{aligned}\mathcal{L}_h(\theta_F,\theta_f,\theta_g,\theta_s) = {} & \mu_f\mathcal{L}_f(y_f, G_f(G(x;\theta_F);\theta_f))+\\ & \mu_g\mathcal{L}_g(y_g, G_g(G(x;\theta_F);\theta_g))+\\ & \mathcal{L}_s(y_s, G_s(G(x;\theta_F);\theta_s))\end{aligned} \tag{1}$$

where $\mathcal{L}_f$, $\mathcal{L}_g$ and $\mathcal{L}_s$ denote the loss of family-level, genus-level, and species-level classification subnetworks, respectively. y_f, y_g and y_s respectively denote the true label of the image at family-level, genus-level, and species level. θ_F denote the parameters of the feature extractor (the first five groups of convolutional layers) in VGG-16Net. θ_f, θ_g, and θ_s denote the parameters of family-level, genus-level, and species-level classification subnetworks. μ_f and μ_g denote weights of loss of family and genus-level classification subnetworks. G and G_f (or G_g, G_s) respectively denote the feature vector of VGG-16Net, family-level (or genus-level, species-level) hierarchical classification subnetworks.

Note that the labels of training data do not include unseen classes and thus domain shift will occur when we extract features for test images using the deep neural networks trained by this training data. To avoid domain shift, we add a domain adaptation structure[38], which includes a gradient reversal layer and a domain classifier, after the fifth group of convolutional layers in VGG-

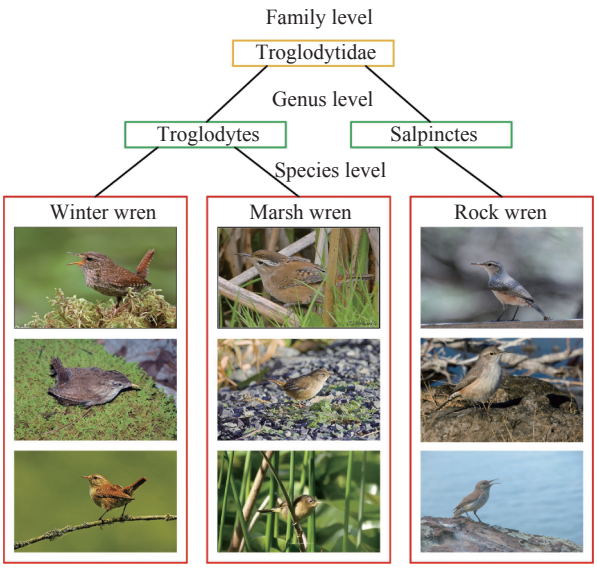


Fig. 2 Hierarchical semantic structure of fine-grained classes

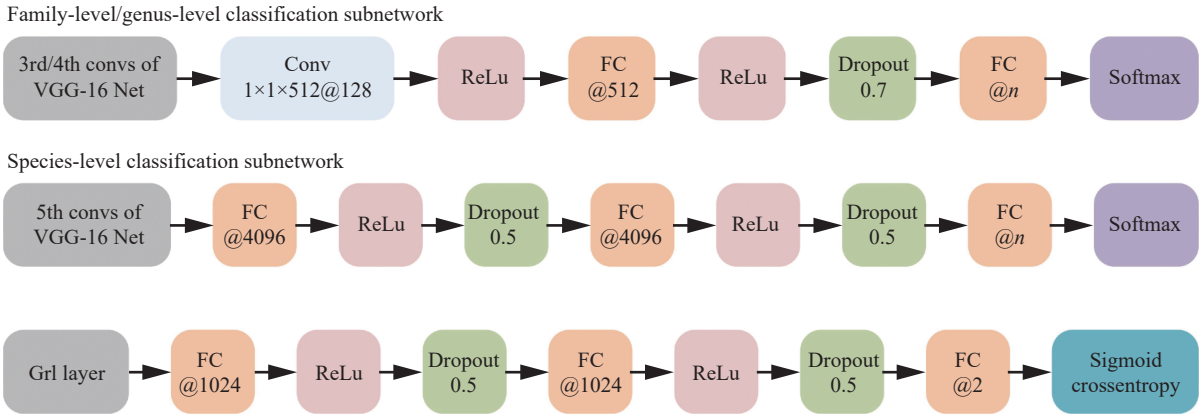


Fig. 3 Detailed architecture of hierarchical classification subnetworks. In this figure, "Conv" and "FC" denote the convolutional layer and fully-connected layer respectively. The numbers under the "Conv", "FC" and "Dropout" denote the kernel information of the convolutional layer, number of output of fully-connected layer and the ratio of dropout, respectively. n is the total number of classes at the corresponding level.

16Net (as shown in Fig. 1). The domain adaption structure views training data and test data as two domains and aims to train a domain classifier that cannot distinguish the domain of some given data. In this way, the difference of features among data from two domains can be eliminated. In our model, we aim to achieve an adversarial process, i.e., to learn features that can confuse the domain classifier and classify fine-grained classes. Therefore, we aim to minimize the loss of hierarchical classification subnetworks and maximize the loss of the domain classifier. The gradient reversal layer (Grl layer in Fig. 4) proposed by Ganin and Lempitsky[38] is used to achieve the goal. We also present the detailed architecture of a domain classifier in Fig. 4. By merging the domain adaptation structure, hierarchical classification subnetworks and VGG-16 Net together, the total loss of an image x is given as follows:

$$\mathcal{L}(\theta_F, \theta_f, \theta_g, \theta_s, \theta_d) = \mathcal{L}_h(\theta_F, \theta_f, \theta_g, \theta_s) - \mu_d \mathcal{L}_d(y_d, G_d(G_s(\theta_s, G(x; \theta_F)); \theta_d)) \quad (2)$$

where $\mathcal{L}_d$, y_d, μ'_d and θ_d respectively denote the loss of domain classifier, the domain label of image x, the weight of loss of domain classifier and the parameter of domain classifier. G_d denotes the domain classifier.

It should be noted that the hierarchical semantic structure of fine-grained classes actually plays an important role in extracting discriminative features for zero-shot fine-grained classification. Figs. 5–6 provide some samples of misclassified images when only using species-level (or species-level/genus-level) features. From Figs. 5–6, we can observe that there are obvious semantic relations between the true labels and predicted labels of these misclassified images (in blue boxes), which are visually similar to images in their predicted classes (in red boxes). Hence, the hierarchical semantic structure of classes can be used to capture discriminative features and thus lead to better recognition results.

3.2 Label inference

In this section, with the discriminative features obtained from Section 3.1, we provide a label propagation approach for zero-shot fine-grained image classification.

We use $S = \{s_1, \cdots, s_p\}$ to represent the set of seen classes, where p is the number of seen classes. And we use $U = \{u_1, \cdots, u_q\}$ to represent the set of unseen classes, where q is the number of unseen classes. Specifically, classes which appeared in S won't appear in U, i.e., $S \cap U = \phi$. We are given a training set of size N_s, denoted as $D_s = \{(x_i, y_i) : i = 1, \cdots, N_s\}$. In the training set, the feature vector of the i-th image in the training set is denoted as x_i, the corresponding label is denoted as $y_i \in S$. We are also given a test set of size N_u, denoted as $D_u = \{(x_j, y_j) : j = 1, \cdots, N_u\}$. In the test set, the feature vector of the j-th image in the test set is denoted as

Fig. 4 Detailed architecture of domain classifier. In this figure, "Grl" and "FC" denote gradient reversal layer and the fully-connected layer, respectively. The numbers under the "FC" and "Dropout" denote the number of output of fully-connected layer and the ratio of dropout, respectively.

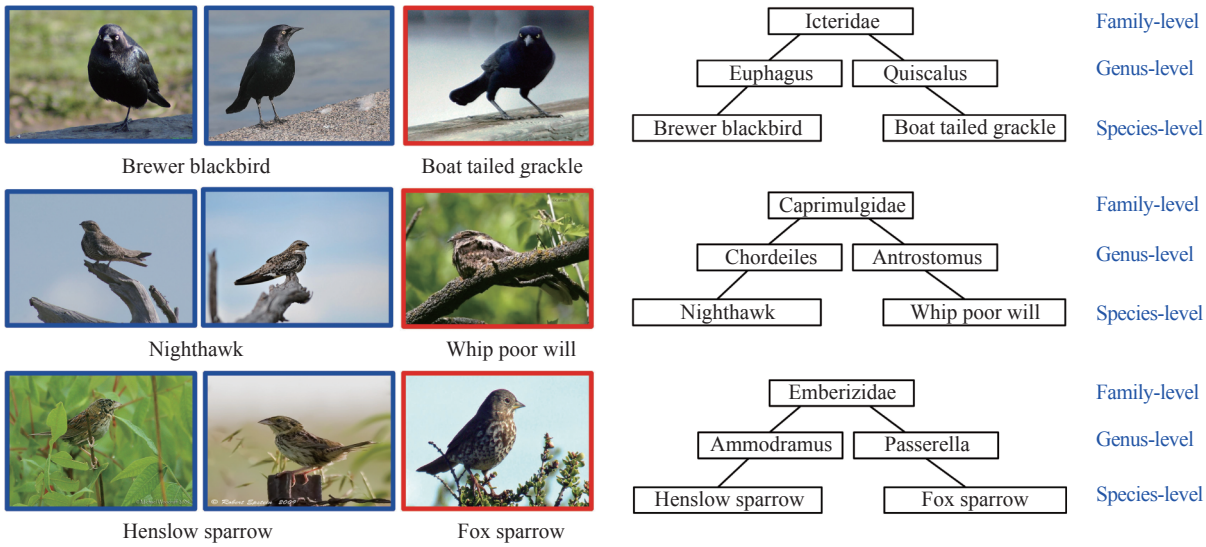


Fig. 5 Samples of misclassification when only using species-level features. In this figure, images in the blue boxes are misclassified images when only using species-level features and the category names under images are their true labels, while the predicted labels of these images (in blue boxes) are given with a sample (in red boxes) in the corresponding rows. These misclassified images are correctly classified when the species-level/genus-level features are used.

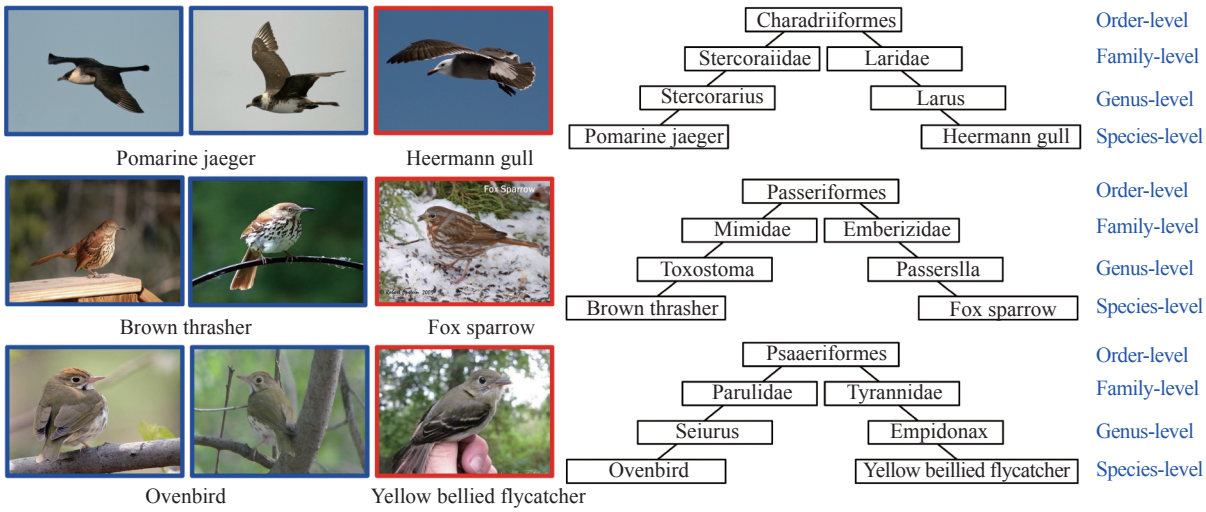


Fig. 6 Samples of misclassification when only using species-level/genus-level features. In this figure, images in the blue boxes are misclassified images when only using species-level/genus-level features and the category names under images are their true labels, while the predicted labels of these images (in blue boxes) are given with a sample (in red boxes) in the corresponding rows. These misclassified images are correctly classified when the species-level/genus-level/family-level features are used.

x_j, the corresponding unknown label is denoted as $y_j \in U$. The main goal of zero-shot learning is to learn a classifier that can predict the correct label y_j for a test image x_j.

In order to predict the labels of images in unseen classes from seen classes, we should measure the semantic relationships between seen and unseen classes at first. In this paper, we collect the attributes of each fine-grained class to form its semantic vector. After that, a semantic-directed graph $\mathcal{G} = \{V, E\}$ is used over all classes (including seen/unseen classes). The set of nodes (i.e., fine-grained classes) in the graph is denoted as V and the set of directed edges between classes is denoted as E Three steps are used to construct the graph $\mathcal{G}$:

1) The first step is to construct the edges between seen classes. A k-nearest-neighbors (k-NN) method is used on semantic vectors for each seen class. We construct a directed edge from a seen class to classes that are k_1 nearest neighbors of it in seen classes. Specifically, the Euclidean distance between two classes is the weight of the edge between two nodes.

2) The second step is to construct the edges between seen classes and unseen classes. A k-nearest-neighbors (k-NN) method is used on semantic vectors for each seen class. We construct a directed edge from a seen class to classes that are k_2 nearest neighbors of it in unseen classes. Specifically, the Euclidean distance between two classes is the weight of the edge between two nodes.

3) Finally, for each unseen classes, it has one edge pointing to itself whose weight is 1.

A weight matrix W of the semantic-directed graph $\mathcal{G}$ is constructed as

$$W = \begin{bmatrix} R_1 & R_2 \\ 0 & I \end{bmatrix} \tag{3}$$

where $R_1 \in \boldsymbol{R}^{p\times p}$ represents the weights among seen classes, $R_2 \in \boldsymbol{R}^{p\times q}$ represents the weights between seen classes and unseen classes, and $I \in \boldsymbol{R}^{q\times q}$ is an identity matrix. Given a weight matrix W, we can define a Markov chain process:

$$T = D^{-1}W \tag{4}$$

where D is a diagonal matrix whose i-th diagonal element is equal to the sum of the i-th row of W.

A normalization method is then exploited to guarantee that the Markov chain process has a unique stationary solution[39, 40]:

$$P = \frac{\eta}{p+q-1}(1_{p+q} - I_{p+q}) + (1-\eta)T \tag{5}$$

where η is a normalization parameter, 1_{p+q} is a one matrix and I_{p+q} is an identity matrix, the size of both of them is $(p+q)\times(p+q)$.

Zero-shot fine-grained classification can be formulated as the following label propagation problem:

$$\min_{\tilde{Y}_i} \frac{1}{2}\sum_{u,v}\pi(u)p_{uv}\left(\frac{\tilde{Y}_{iu}}{\sqrt{\pi(u)}} - \frac{\tilde{Y}_{iv}}{\sqrt{\pi(v)}}\right)^2 + \lambda\|\tilde{Y}_i - Y_i\|_2^2 \tag{6}$$

where $\tilde{Y}_i$ (the i-th row of $\tilde{Y} \in \boldsymbol{R}^{N_u\times(p+q)}$) is the optimal probabilities of the i-th test image belonging to each class. Y_i (the i-th row of $Y \in \boldsymbol{R}^{N_u\times(p+q)}$) is the initial probabilities of the i-th test image belonging to each class. Y is an initialization of $\tilde{Y}$ and $\tilde{Y}$ is the final solution of the problem formulated in (6). Moreover, $\pi(u)$ is the sum of the u-th row of P (i.e., $\sum_v p_{uv}$), and λ is a regularization parameter.

In order to ensure that semantically similar classes for the i-th test image have similar $\tilde{Y}_i$, the first term of the above objective function sums the weighted variation of $\tilde{Y}_i$ on each edge of the directed graph $\mathcal{G}$. In order to en-

sure that $\tilde{Y}_{i\cdot}$ does not change too much from Y_i, the second term denotes an L_2-norm fitting constraint.

We adopt the technique introduced by Zhou et al.[39] in order to solve the above label propagation problem. In this paper, we define the operator Θ:

$$\Theta = \frac{(\Pi^{\frac{1}{2}} P \Pi^{-\frac{1}{2}} + \Pi^{-\frac{1}{2}} P \Pi^{\frac{1}{2}})}{2} \tag{7}$$

where Π is a diagonal matrix with size $(p+q) \times (p+q)$. And the u-th diagonal element of the matrix is equal to $\pi(u)$. The optimal solution $\tilde{Y}^*$ of the problem in (6) is

$$\tilde{Y}^* = Y(I - \alpha\Theta)^{-1} \tag{8}$$

according to Zhou et al.[39], where $I \in \boldsymbol{R}^{(p+q)\times(p+q)}$ is an identity matrix and $\alpha = \dfrac{1}{(1+\lambda)} \in (0,1)$.

Y should be provided in advance in order to obtain the above solution. There are two parts in each row of Y: the probabilities of a test image belonging to seen classes, and the probabilities of a test image belonging to unseen classes. We set the probabilities belonging to unseen classes as 0 since there is no labelled data in unseen classes. We use LIBLINEAR toolbox[41] to train an L_2-regularized logistic regression classifier, in order to compute the initial probabilities belonging to seen classes. In general, we empirically set the parameter c in L_2-regularized logistic regression as 0.01.

To sum up, by combining the feature learning and label propagation approaches together, the complete algorithm for zero-shot fine-grained classification is outlined as Algorithm 1. It should be noted that the proposed approach can be extended to a weak supervision setting by replacing class attributes with semantic vectors extracted by word vector extraction methods (e.g., Word2Vec[16]).

Algorithm 1. The proposed framework

Input: the set of labeled training images D_s, the set of test images in unseen classes X_u

Feature learning:

1) Train the proposed neural network using hierarchical semantic structure among fine-grained classes;

2) Run forward computation of the proposed neural network for each test image and extract deep features from hierarchical classification subnetworks;

3) Concatenate the features from hierarchical classification subnetworks to obtain deep features F;

Label inference:

4) Compute the initial probabilities of test images belonging to unseen classes Y with the LIBLINEAR toolbox[41] and deep features F;

5) Construct the semantic-directed graph based on semantic vectors;

6) Compute the normalized transition matrix P according to (3)–(5);

7) Find the solution $\tilde{Y}^*$ of label propagation problem formulated in (6) according to (7) and (8);

8) Label each test image x_i with class $\arg\max_j \tilde{Y}^*_{ij}$.

Output: Labels of test images in unseen classes.

4 Experimental results

4.1 Experimental setup

In our experiment, we describe our experiments on the benchmark fine-grained datasets, Caltech UCSD Birds-200-2011[6] and Oxford Flower-102[10].

4.1.1 Caltech UCSD Birds-200-2011 dataset

The Caltech UCSD Birds-200-2011 dataset (6) contains 11, 788 images of 200 North-American bird species[6]. Each species is associated with a Wikipedia article and organized by scientific classification (family, genus, species). Each class is also annotated with 312 visual attributes. In the zero-shot setting, we follow Zhang and Saligrama[33] to use 150 bird species as seen classes for training and the left 50 species as unseen classes for testing. The results are the average of four fold cross validation. For parameter validation, we also use a zero-shot setting within the 150 classes of the training set, i.e., we use 100 classes for training and the rest for validation. The hierarchical labels of fine-grained classes are collected from Wikipedia. For each fine-grained class, we use 312-d class attributes and 300-d semantic vectors extracted by the 16 model[16] (trained by GoogleNews) as semantic description.

4.1.2 Oxford Flower-102 dataset

The Oxford Flower-102 (Flowers-102) dataset contains 8 189 images of 102 different categories. There is no human annotated attribute for each category. Therefore, we choose 80 of 102 categories, which are associated with a Wikipedia article and organized by scientific classification (family, genus, species). In the zero-shot setting, similar to the setting of 6 dataset, we use 60 flower species as seen classes for training and the left 20 species as unseen classes for testing. The results are the average of four fold cross validation. For parameter validation, we use a similar strategy as the 6 dataset, i.e., we use 60 classes for training and the rest for validation. The hierarchical labels of fine-grained classes are collected from Wikipedia. For each fine-grained class, only 300-d semantic vectors extracted by the 16 model[16] (trained by GoogleNews) are used as semantic description.

4.2 Implementation details

In the feature learning phase, the VGG-16′s layers are pre-trained on ILSVRC 2012 1K classification[12], and then finetuned with training data. Meanwhile, other layers are trained from scratch. All input images are resized to 224×224. Stochastic gradient descent (SGD)[42] is used to optimize our model with a basic learning rate of 0.01, a momentum of 0.9, a weight decay of 0.005 and a mini-batch size of 20. For layers trained from scratch, their

learning rate is 10 times that of the basic learning rate. The model is implemented based on Caffe[43]. In the label inference phase, we choose the parameters μ_f, μ_g, η and λ by cross validation on training data.

In this phase, different-level features are extracted from the last but one fully-connected layers before softmax layers and we finally obtain three kinds of features which are used to classify images at different levels. To find a good way to combine these features, we conduct experiments on the proposed model using the concatenation of different-level features and the results are shown in Table 1. From the table, we can observe that high-level features perform better than features extracted from shallow layers. Furthermore, the combination of three-level features performs best. Therefore, we use the concatenation of three-level features as the final deep visual features in our model.

4.3 Effectiveness of the proposed feature learning approach

To test the effectiveness of the proposed feature learning approach, we utilize features extracted by the the proposed feature learning approach into other zero-shot learning models[13–15] and results are given in Fig. 7. From Fig. 7, we can observe that the proposed feature learning method works well in other zero-shot learning models. Compared with features extracted from VGG-16Net pretrained by ImageNet, the proposed feature learning approach involves hierarchical semantic structure of labels and domain adaptation structure, which thus generate more discriminative features for zero-shot fine-grained classification.

4.4 Comparison with state-of-the-arts

4.4.1 Testing on class attributes

We provide the comparison of the proposed approach to the state-of-the-art zero-shot learning approaches[13–15, 33–35] using class attributes on the 6 dataset, which is shown in Table 2. In this table, "ZC" denotes the zero-shot learning approach based on label propagation, "VGG-16Net" denotes the features obtained from VGG-16Net[37] (pretrained with ImageNet[12]), "HCS" denotes the hierarchic-

Table 1 Comparison of the proposed approach using the concatenation of different-level features on the CUB–200–2011 dataset

Features	Semantic level	Accuracy (%)	
		Class attributes	Semantic vectors
Finetuned VGG-16Net+HCS	Species-level	44.9	28.9
	Genus-level	36.3	22.3
	Family-level	32.3	15.3
	Species-level/Genus-level	45.7	30.4
	Species-level/Genus-level/Family-level	**46.2**	**32.2**
Finetuned VGG-16Net+HCS+DA	Species-level	46.8	29.8
	Genus-level	37.1	24.3
	Family-level	33.2	18.3
	Species-level/Genus-level	48.3	33.2
	Species-level/Genus-level/Family-level	**49.5**	**34.5**

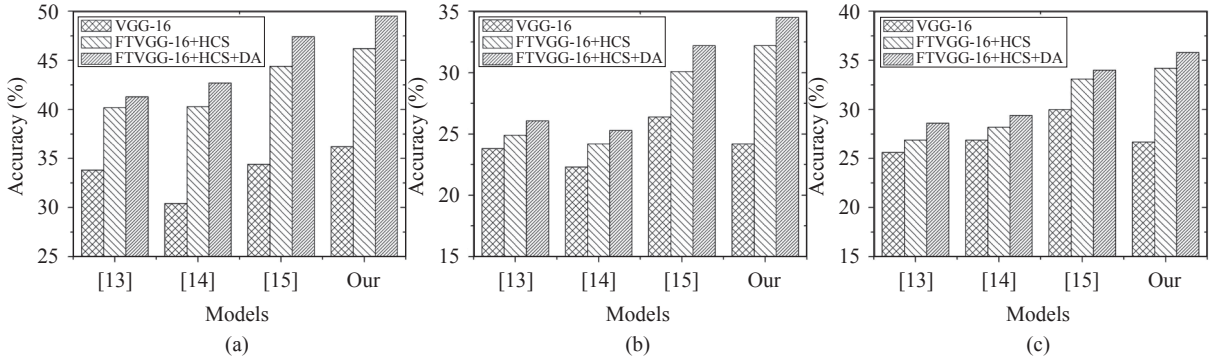


Fig. 7 The results of other zero-shot learning approaches using the proposed feature learning approach. (a) 6 dataset using class attributes. (b) 6 dataset using semantic vectors. (c) Flowers-102 dataset using semantic vectors. In this figure, "VGG-16", "FTVGG-16+HCS" and " FTVGG-16+HCS+DA" denote features obtained from VGG-16Net pretrained by ImageNet, VGG-16Net with hierarchical classification subnetworks and VGG-16Net with hierarchical classification subnetworks and domain adaptation structure, respectively.

al classification subnetworks of the method proposed in Section 3.1, and "DA" denotes domain adaptation structure of the method in Section 3.1. It can be seen that the proposed approach significantly outperforms the state-of-art zero-shot learning approaches. The comparison between "ZC" versus "Our full model" demonstrates that the proposed feature learning approach is very effective in the task of zero-shot fine-grained classification. The comparison between "ZSC" versus "Our full model" demonstrates that the domain adaptation structure is necessary for feature learning in the task of zero-shot fine-grained classification. It should be noted that Akata et al.[35] has achieved 56.5% using a multi-cue framework, where locations of parts are involved as very strong supervision in the training and test process. The 49.5% released in Table 2 is its classification result when annotations of the whole images are used (without locations of parts). The superior performance of the proposed approach compared with Akata et al.[35] verifies the effectiveness of the proposed approach.

4.4.2 Testing on semantic vectors

We also evaluate the proposed approach in the weakly supervised setting, where only fine-grained labels of training images are given and the semantics among fine-grained are learned from text descriptions. Table 3 provides classification results on 6 and Flowers-102 datasets in the weaker supervised setting. From this table, we can observe that our approach outperforms the state-of-the-art zero-shot learning algorithms, which verifies the effectiveness of the proposed model.

5 Conclusions

In this paper, we propose a two-phase framework for zero-shot fine-grained classification approach, which can recognize images from unseen fine-grained classes. In our approach, a feature learning strategy based on hierarchical semantic structures of fine-grained classes and class attributes is developed to generate robust and discriminative features and then a label propagation method based

Table 2 Comparison of zero-shot learning approaches on class attributes

Datasets	Approaches	Features	Accuracy (%)
CUB-200-2011	[13]	VGG-16Net	33.8
	[14]	VGG-16Net	30.4
	[15]	VGG-16Net	34.4
	[33]	VGG-16Net	42.1
	[34]	VGG-16Net	43.3
	[35]	VGG-16Net	43.3
	ZC	VGG-16Net	36.2
	ZSC	Finetuned VGG-16Net+HCS	46.2
	Our full model	Finetuned VGG-16Net+HCS+DA	**49.5**
Flowers-102	–	–	–

Table 3 Comparison of zero-shot learning approaches on the semantic vectors

Datasets	Approaches	Features	Accuracy (%)
CUB-200-2011	[13]	VGG-16Net	23.8
	[14]	VGG-16Net	22.3
	[15]	VGG-16Net	26.4
	[36]	VGG-16Net	29.0
	ZC	VGG-16Net	24.2
	ZSC	Finetuned VGG-16Net+HCS	32.2
	Our full model	Finetuned VGG-16Net+HCS+DA	**34.5**
Flowers-102	[13]	VGG-16Net	25.6
	[14]	VGG-16Net	27.3
	[15]	VGG-16Net	30.8
	ZC	VGG-16Net	26.7
	ZSC	Finetuned VGG-16Net+HCS	34.2
	Our full model	Finetuned VGG-16Net+HCS+DA	**35.8**

on semantic directed graphs is proposed for label inference. Experimental results on the benchmark fine-grained classification datasets demonstrate that the proposed approach outperforms state-of-the-art zero-shot learning algorithms. Our approach can be extended to the weakly supervised setting (i.e., only fine-grained labels of training images are given) and has achieved better results than the state-of-the-art. In future work, we will make further improvements on developing more powerful word vector extractors to explore better semantic relationships among fine-grained classes and optimize the feature extractors with word vector extractors simultaneously.

Acknowledgement

This work was supported by National Basic Research Program of China (973 Program) (No. 2015CB352502), National Nature Science Foundation of China (No. 61573026) and Beijing Nature Science Foundation (No. L172037).

Open Access

References

[1] B. Zhao, J. S. Feng, X. Wu, S. C. Yan. A survey on deep learning-based fine-grained object classification and semantic segmentation. *International Journal of Automation and Computing*, vol. 14, no. 2, pp. 119–135, 2017. DOI: 10.1007/s11633-017-1053-3.

[2] M. El Mallahi, A. Zouhri, A. El Affar, A. Tahiri, H. Qjidaa. Radial Hahn moment invariants for 2D and 3D image recognition. *International Journal of Automation and Computing*, vol. 15, no. 3, pp. 277–289, 2018. DOI: 10.1007/s11633-017-1071-1.

[3] H. S. Du, Q. P. Hu, D. F. Qiao, I. Pitas. Robust face recognition via low-rank sparse representation-based classification. *International Journal of Automation and Computing*, vol. 12, no. 6, pp. 579–587, 2015. DOI: 10.1007/s11633-015-0901-2.

[4] T. Long, X. Xu, F. M. Shen, L. Liu, N. Xie, Y. Yang. Zero-shot learning via discriminative representation extraction. *Pattern Recognition Letters*, vol. 109, pp. 27–34, 2018. DOI: 10.1016/j.patrec.2017.09.030.

[5] E. Kodirov, T. Xiang, S. G. Gong. Semantic autoencoder for zero-shot learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 3174–3183, 2017. DOI: 10.1109/CVPR.2017.473.

[6] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset, Technical Report CNS-TR-2011-001, California Institute of Technology, USA, 2011.

[7] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, P. Perona. Caltech-UCSD Birds 200, Technical Report CNS-TR-2010-001, California Institute of Technology, USA, 2010.

[8] T. Berg, J. X. Liu, S. W. Lee, M. L. Alexander, D. W. Jacobs, P. N. Belhumeur. Birdsnap: Large-scale fine-grained visual categorization of birds. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, USA, pp. 2019–2026, 2014. DOI: 10.1109/CVPR.2014.259.

[9] B. P. Yao, A. Khosla, F. F. Li. Combining randomization and discrimination for fine-grained image categorization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Springs, Colorado, USA, pp. 1577–1584, 2011.

[10] M. E. Nilsback, A. Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the 6th Indian Conference on Computer Vision, Graphics & Image Processing*, IEEE, Bhubaneswar, India, pp. 722–729, 2008. DOI: 10.1109/ICVGIP.2008.47.

[11] A. R. Sfar, N. Boujemaa, D. Geman. Vantage feature frames for fine-grained categorization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Portland, USA, pp. 835–842, 2013. DOI: 10.1109/CVPR.2013.113.

[12] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. H. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, L. Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015. DOI: 10.1007/s11263-015-0816-y.

[13] B. Romera-Paredes, P. H. S. Torr. An embarrassingly simple approach to zero-shot learning. In *Proceedings of the 32nd International Conference on Machine Learning*, ACM, Lille, France, pp. 2152–2161, 2015.

[14] Z. M. Zhang, V. Saligrama. Zero-shot learning via semantic similarity embedding. In *Proceedings of IEEE International Conference on Computer Vision*, Santiago, Chile, pp. 4166–4174, 2015. DOI: 10.1109/ICCV.2015.474.

[15] Z. Y. Fu, T. A. Xiang, E. Kodirov, S. G. Gong. Zero-shot object recognition by semantic manifold distance. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA, pp. 2635–2644, 2015. DOI: 10.1109/CVPR.2015.7298879.

[16] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, Curran Associates Inc., Lake Tahoe, USA, pp. 1188–1196, 2013.

[17] N. Zhang, J. Donahue, R. Girshick, T. Darrell. Part-based R-CNNs for fine-grained category detection. In *Proceedings of the 13th European Conference on Computer Vision*, Springer, Zurich, Switzerland, pp. 834–849, 2014. DOI: 10.1007/978-3-319-10590-1_54.

[18] S. L. Huang, Z. Xu, D. C. Tao, Y. Zhang. Part-stacked CNN for fine-grained visual categorization. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 1173–1182, 2016. DOI: 10.1109/CVPR.2016.132.

[19] A. Krizhevsky, I. Sutskever, G. E. Hinton. ImageNet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, Curran Associates Inc., Lake Tahoe, USA, pp. 1097–1105, 2012.

[20] Z. Xu, S. L. Huang, Y. Zhang, D. C. Tao. Augmenting strong supervision using web data for fine-grained categorization. In *Proceedings of IEEE International Conference on Computer Vision*, Santiago, Chile, pp. 2524–2532, 2015. DOI: 10.1109/ICCV.2015.290.

[21] H. Zhang, T. Xu, M. Elhoseiny, X. L. Huang, S. T. Zhang, A. Elgammal, D. Metaxas. SPDA-CNN: Unifying semantic part detection and abstraction for fine-grained recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 1143–1152, 2016. DOI: 10.1109/CVPR.2016.129.

[22] C. Szegedy, W. Liu, Y. Q. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich. Going deeper with convolutions. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Boston, USA, pp. 1–9, 2015. DOI: 10.1109/CVPR.2015.7298594.

[23] T. Y. Lin, A. RoyChowdhury, S. Maji. Bilinear CNN models for fine-grained visual recognition. In *Proceedings of IEEE International Conference on Computer Vision*, Santiago, Chile, pp. 1449–1457, 2015. DOI: 10.1109/ICCV.2015.170.

[24] X. P. Zhang, H. K. Xiong, W. G. Zhou, W. Y. Lin, Q. Tian. Picking deep filter responses for fine-grained image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 1134–1142, 2016. DOI: 10.1109/CVPR.2016.128.

[25] E. Kodirov, T. Xiang, Z. Y. Fu, S. G. Gong. Unsupervised domain adaptation for zero-shot learning. In *Proceedings of IEEE International Conference on Computer Vision*, Santiago, Chile, pp. 2452–2460, 2015. DOI: 10.1109/ICCV.2015.282.

[26] C. H. Lampert, H. Nickisch, S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 3, pp. 453–465, 2014. DOI: 10.1109/TPAMI.2013.140.

[27] P. Kankuekul, A. Kawewong, S. Tangruamsub, O. Hasegawa. Online incremental attribute-based zero-shot learning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Providence, USA, pp. 3657–3664, 2012. DOI: 10.1109/CVPR.2012.6248112.

[28] M. Rohrbach, M. Stark, B. Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Springs, Colorado USA, pp. 1641–1648, 2011. DOI: 10.1109/CVPR.2011.5995627.

[29] X. D. Yu, Y. Aloimonos. Attribute-based transfer learning for object categorization with zero/one training example. In *Proceedings of the 11th European Conference on Computer Vision*, Springer, Heraklion, Greece, pp. 127–140, 2010. DOI: 10.1007/978-3-642-15555-0_10.

[30] M. Palatucci, D. Pomerleau, G. Hinton, T. M. Mitchell. Zero-shot learning with semantic output codes. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems*, Curran Associates Inc., Vancouver, Canada, pp. 1410–1418, 2009.

[31] C. H. Lampert, H. Nickisch, S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Miami, USA, pp. 951–958, 2009. DOI: 10.1109/CVPR.2009.5206594.

[32] Y. Q. Xian, B. Schiele, Z. Akata. Zero-shot learning-the good, the bad and the ugly. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, USA, pp. 4582–4591, 2017. DOI: 10.1109/CVPR.2017.328.

[33] Z. M. Zhang, V. Saligrama. Zero-shot learning via joint latent similarity embedding. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 6034–6042, 2016. DOI: 10.1109/CVPR.2016.649.

[34] M. Bucher, S. Herbin, F. Jurie. Improving semantic embedding consistency by metric learning for zero-shot classification. In *Proceedings of the 14th European Conference on Computer Vision*, Springer, Amsterdam, The Netherlands, pp. 730–746, 2016. DOI: 10.1007/978-3-319-46454-1_44.

[35] Z. Akata, M. Malinowski, M. Fritz, B. Schiele. Multi-cue zero-shot learning with strong supervision. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Las Vegas, USA, pp. 59–68, 2016. DOI: 10.1109/CVPR.2016.14.

[36] R. Z. Qiao, L. Q. Liu, C. H. Shen, A. van den Hengel. Less is more: Zero-shot learning from online textual documents with noise suppression. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 2441–2448, 2016. DOI: 10.1109/CVPR.2016.247.

[37] K. Simonyan, A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of International Conference on Learning Representations*, ICLR, San Diego, USA, pp. 59–68, 2015.

[38] Y. Ganin, V. S. Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, ICML, Lille, France, pp. 1180–1189, 2015.

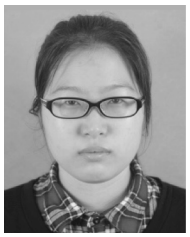
[39] D. Y. Zhou, J. Y. Huang, B. Scholkopf. Learning from labelled and unlabelled data on a directed graph. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML, Bonn, Germany, pp. 1036–1043, 2005.

[40] A. X. Li, Z. W. Lu, L. W. Wang, T. Xiang, J. R. Wen. Zero-shot scene classification for high spatial resolution remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 4157–4167, 2017. DOI: 10.1109/TGRS.2017.2689071.

[41] R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang, C. J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[42] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989. DOI: 10.1162/neco.1989.1.4.541.

[43] Y. Q. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM International Conference on Multimedia*, ACM, Orlando, USA, 2014. DOI: 10.1145/2647868.2654889.



Ao-Xue Li received the B. Sc. degree in electronic science and technology from Beijing Normal University, China in 2015. She is currently a Ph. D. degree candidate in computer science and technology at Peking University, China.

Her research interests include computer vision and machine learning.

E-mail: lax@pku.edu.cn (Corresponding author)

ORCID iD: 0000-0002-5467-9405

Ke-Xin Zhang received the B. Sc. degree in computer science and technology from Peking University, China in 2018. She is currently a master stadent in computer science and technology at Peking University, China.

Her research interests include computer vision and machine learning.

E-mail: zhangkexin@pku.edu.cn

Li-Wei Wang received the B. Sc. and M. Sc. degrees in electronic engineering from Department of Electronic Engineering, Tsinghua University, China in 1999 and 2002, respectively, the Ph. D. degree in applied mathematics from School of Mathematical Sciences, Peking University, China in 2005. He is currently a full professor of School of Electronics Engineering and Computer Sciences, Peking University, China. He has published about 100 refereed journal and conference papers. He was named among "AI′s 10 to Watch" in 2010.

His research interest is machine learning, with application to computer vision.

E-mail: wanglw@cis.pku.edu.cn