

Object-Part Attention Model for Fine-Grained Image Classification

Yuxin Peng¹, Xiangteng He, and Junjie Zhao

Abstract—Fine-grained image classification is to recognize hundreds of subcategories belonging to the same basic-level category, such as 200 subcategories belonging to the bird, which is highly challenging due to *large* variance in the same subcategory and *small* variance among different subcategories. Existing methods generally first locate the objects or parts and then discriminate which subcategory the image belongs to. However, they mainly have *two limitations*: 1) relying on object or part annotations which are heavily labor consuming; and 2) ignoring the spatial relationships between the object and its parts as well as among these parts, both of which are significantly helpful for finding discriminative parts. Therefore, this paper proposes the *object-part attention model (OPAM)* for weakly supervised fine-grained image classification and the main novelties are: 1) *object-part attention model* integrates two level attentions: *object-level attention* localizes objects of images, and *part-level attention* selects discriminative parts of object. Both are jointly employed to learn multi-view and multi-scale features to enhance their mutual promotion; and 2) *Object-part spatial constraint model* combines two spatial constraints: *object spatial constraint* ensures selected parts highly representative and *part spatial constraint* eliminates redundancy and enhances discrimination of selected parts. Both are jointly employed to exploit the subtle and local differences for distinguishing the subcategories. Importantly, neither object nor part annotations are used in our proposed approach, which avoids the heavy labor consumption of labeling. Compared with more than ten state-of-the-art methods on four widely-used datasets, our OPAM approach achieves the best performance.

Index Terms—Fine-grained image classification, object-part attention model, object-part spatial constraint model, weakly supervised learning.

I. INTRODUCTION

FINE-GRAINED image classification is highly challenging, aiming to recognize hundreds of subcategories under the same basic-level category, such as hundreds of subcategories of birds [1], cars [2], pets [3], flowers [4] and aircrafts [5]. While basic-level image classification only needs to discriminate the basic-level category, such as bird or car. The difference between basic-level and fine-grained image classification is shown as Fig. 1. Fine-grained image classification is a highly important task with wide applications,

Manuscript received April 6, 2017; revised September 20, 2017; accepted November 7, 2017. Date of publication November 15, 2017; date of current version December 27, 2017. This work was supported by the National Natural Science Foundation of China under Grant 61771025, Grant 61371128, and Grant 61532005. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Guoliang Fan. (Corresponding author: Yuxin Peng.)

The authors are with the Institute of Computer Science and Technology, Peking University, Beijing 100871, China (e-mail: pengyuxin@pku.edu.cn).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TIP.2017.2774041





Inputs				
Basic-level Image Classification	Bird	Bird	Car	Car
Fine-grained Image Classification	American Crow	Fish Crow	Hyundai Elantra Sedan 2007	Toyota Sequoia SUV 2012

Fig. 1. Basic-level image classification vs. fine-grained image classification. In basic-level image classification, we only need to classify the first two images as bird category, distinguishing them from car category. While in fine-grained image classification, the subcategory should be further determined exactly. For example, the first two images belong to the subcategories of American Crow and Fish Crow respectively.

such as automatic driving, biological conservation and cancer detection. Fig. 2 shows the large variance in the same subcategory and small variance among different subcategories. It is extremely hard for human beings to recognize hundreds of subcategories, such as 200 bird subcategories or 196 car subcategories. Due to small variance in object appearances, subtle and local differences are the key points for fine-grained image classification, such as the color of back, the shape of bill and the texture of feather for bird. These subtle and local differences locate at the discriminative objects and parts, most existing methods [6]–[8] generally follow the strategy of locating the objects or parts in the image and then discriminating which subcategory the image belongs to.

To localize the discriminative objects and parts, generating image patches with high objectness by a bottom-up process is generally first performed, meaning that the generated patches contain the discriminative object or parts. Selective search [9] is an unsupervised method that can generate thousands of such image patches, which is extensively used in recent methods [6], [7], [10]. Since the bottom-up process has high recall but low precision, it is indispensable to remove the noisy image patches and retain those containing the object or discriminative parts, which can be achieved through top-down attention model. In the context of fine-grained image classification, finding the objects and discriminative parts can be regarded as a two-level attention process, where one is object-level and the other is part-level. An intuitive idea is to use object annotation (i.e. bounding box of object) for object-level attention and part annotations (i.e. part locations) for part-level attention, such as [6], [11]–[13], but the labeling is heavily labor consuming. This is the *first limitation*.

For addressing the above problem, researchers begin focusing on how to achieve promising performance under



Fig. 2. Illustration of challenges in fine-grained image classification: large variance in same subcategory as shown in the first row, and small variance among different subcategories as shown in the second row. The images in (a) Birds, (b) Cars, (c) Cats and (d) Flowers are from CUB-200-2011 [1], Cars-196 [2], Oxford-IIIT Pet [3] and Oxford-Flower-102 [4] datasets respectively.

the weakly supervised setting that neither object nor part annotations are used in both training and testing phases. Zhang *et al.* [14] propose to select the discriminative parts through exploiting the useful information in part clusters. Zhang *et al.* [7] propose an automatic fine-grained image classification method, incorporating deep convolutional filters for both part selection and description. However, when they select the discriminative parts, the spatial relationships between the object and its parts as well as among these parts are ignored, but both of them are highly helpful for finding the discriminative parts. This causes the selected parts: (1) have large areas of background noise and small areas of object, (2) have large overlap with each other which leads to redundant information. This is the *second limitation*.

For addressing the above two limitations, this paper proposes the object-part attention model (OPAM) for weakly supervised fine-grained image classification. Its main novelties and contributions can be summarized as follows:

- **Object-Part Attention Model.** Most existing works rely on object or part annotations [6], [12], [13], while labeling is heavily labor consuming. For addressing this important problem, we propose the object-part attention model for weakly supervised fine-grained image classification to avoid using the object and part annotations and march toward practical applications. It integrates two level attentions: (1) **Object-level attention model** utilizes the global average pooling in CNN to extract the saliency map for localizing objects of images, which is to learn object features. (2) **Part-level attention model** first selects the discriminative parts and then aligns the parts based on the cluster patterns of neural network, which is to learn subtle and local features. The object-level attention model focuses on the representative object appearance, and the part-level attention model focuses on the distinguishing specific differences of parts among subcategories. Both of them are jointly employed to boost the multi-view and multi-scale feature learning, and enhance their mutual promotion to achieve good performance for fine-grained image classification.
- **Object-Part Spatial Constraint Model.** Most existing weakly supervised methods [7], [14] ignore the spatial relationships between the object and its parts as well as among these parts, both of which are highly helpful for discriminative part selection. For addressing this problem, we propose the part selection approach

driven by object-part spatial constraint model, which combines two types of spatial constraints: (1) **Object spatial constraint** enforces that the selected parts are located in the object region and highly representative. (2) **Part spatial constraint** reduces the overlaps among parts and highlights the saliency of parts, which eliminates the redundancy and enhances the discrimination of selected parts. Combination of the two spatial constraints not only significantly promotes discriminative part selection by exploiting subtle and local distinction, but also achieves a notable improvement on fine-grained image classification.

Our previous conference paper [15] integrates two level attentions: object-level attention selects image patches relevant to the object, and part-level attention selects discriminative parts, which is the first work to classify fine-grained images without using object and part annotations in both training and testing phases, and achieves promising results [14]. In this paper, our OPAM approach further exploits the two level attentions to localize not only the discriminative parts but also the objects, and employs the object-part spatial constraint model to eliminate redundancy as well as highlight discrimination of the selected parts: **For object-level attention**, we further propose *an automatic object localization approach via saliency extraction* to focus on the representative object feature for better classification performance. It utilizes the global average pooling in CNN for localizing objects of images, rather than only selecting the image patches relevant to object that have large areas of background noise or do not contain the whole object in image like [15]. **For part-level attention**, we further propose *a part selection approach driven by object-part spatial constraint model* to exploit the subtle and local differences among subcategories. It considers the spatial relationships between object and its parts as well as among these parts, thus avoids the problem of generating large areas of background noise and large overlaps among selected parts like [15]. Compared with more than ten state-of-the-art methods on four widely-used datasets, the effectiveness of our OPAM approach is verified by the comprehensive experimental results.

The rest of this paper is organized as follows: Section II briefly reviews related works on fine-grained image classification. Section III presents our proposed OPAM approach, and Section IV introduces the experiments as well as the result analyses. Finally Section V concludes this paper.

II. RELATED WORK

Most traditional methods for fine-grained image classification follow the strategy of extracting basic low-level descriptors like SIFT [16], and then generating Bag-of-Words for image representation [17], [18]. However, the performance of these methods is limited by the handcrafted features. Deep learning has shown its strong power in feature learning, and achieved great progresses in fine-grained image classification [6]–[8], [11], [15], [19]–[25]. These methods can be divided into three groups [26]: ensemble of networks based methods, visual attention based methods and part detection based methods.

A. Ensemble of Networks Based Methods

Ensemble of networks based methods are proposed to utilize multiple neural networks to learn different representations of image for better classification performance. Each subcategory has an implied hierarchy of labels in its ontology tree. For example, *Picoides Pubescens*, which is the label in species level, has the label in genus level as *Picoides* and the family level as *Picidae*. Wang *et al.* [24] first leverage the labels of multiple levels to train a series of CNNs at each level, which focuses on different regions of interest in images. Different features are extracted by different level CNNs, and combined to encode informative and discriminative features. Finally, a linear SVM is trained to learn weights for the final classification. However, the external labels of ontology tree are necessary for the method of [24]. Lin *et al.* [25] propose a bilinear CNN model, which is an end-to-end system jointly combining two CNNs, each of which is adopted as a feature extractor. The extracted features from two CNNs at each location of image are multiplied by outer product operation, and then pooled to generate an image descriptor. Finally, softmax is conducted for final prediction. Despite achieving promising results, these methods are still limited by the lack of ability to be spatially invariant to the input image. Therefore, Jaderberg *et al.* [21] propose a learnable network, called spatial transformer, which consists of three parts: localization network, grid generator and sampler. Four spatial transformers in parallel are performed on images, and capture the discriminative parts to pass to the part description subnets. Finally, softmax is conducted on the concatenated part descriptor for final prediction.

B. Visual Attention Based Methods

Due to attention system, humans focus on the discriminative regions of an image dynamically, rather than receiving and dealing with the information of entire image directly. This natural advantage makes the attention mechanism widely used in fine-grained image classification. Inspired by the way how humans perform visual sequence recognition, Sermanet *et al.* [27] propose the attention for fine-grained categorization (AFGC) system. First, they process a multi-resolution crop on the input image, where each crop is called a glimpse. Then they use the information of glimpses to output the next location and the next object via a deep recurrent neural

network at each step. The final prediction is computed through the sequence of glimpses. Recently, fully convolutional neural network is used to learn the saliency of an image for finding the discriminative regions [28]. Liu *et al.* [29] use the fully convolutional attention to localize multiple parts to get better classification performance. Xie *et al.* [30] propose a novel algorithm, called InterActive, which computes the activeness (attention) of neurons and network connections, carrying high-level context as well as improving the descriptive power of low-level and mid-level neurons, thus achieves good performance on image classification. Zhou *et al.* [28] use global average pooling (GAP) in CNN to generate the saliency map for each image. Based on the saliency map, the discriminative region can be found. Furthermore, a diversified visual attention network (DVAN) [31] is proposed to pursue the diversity of attention as well as gather discriminative information. In this paper, our OPAM approach integrates two level attention models: object-level attention model focuses on the representative object appearance, and part-level attention model focuses on the discriminative parts. Both of them are jointly employed to learn multi-view and multi-scale features to enhance their mutual promotion.

C. Part Detection Based Methods

In fine-grained image classification, subtle and local differences generally locate at discriminative parts of object, so the discriminative part detection is crucial for fine-grained image classification. Girshick *et al.* [10] propose a popular detection method, R-CNN, which first generates thousands of candidate image patches for each image via the bottom-up process [9], and then selects the image patches with high classification scores as detection results. Zhang *et al.* [6] utilize R-CNN with a geometric prior to detect discriminative parts for fine-grained image classification, and then train a classifier on the features of detected parts for final categorization. They use both the object and part annotations.

Recently, researchers begin focusing on how to detect the discriminative parts under the weakly supervised setting, which means neither object nor part annotations are used in both training and testing phases. Simon and Rodner [20] propose a constellation model to localize parts of object, leveraging CNN to find the constellations of neural activation patterns. First, neural activation maps are computed as part detectors by using the outputs of a middle layer of CNN. Second, a part model is estimated by selecting part detectors via constellation model. Finally, the part model is used to extract features for classification. Zhang *et al.* [7] propose an automatic fine-grained image classification method, incorporating deep convolutional filters for both part selection and description. They combine two steps of deep filter response picking: The first step picks the discriminative filters that significantly respond to specific parts in image. The second step picks the salient regions and generates features with spatially weighted Fisher Vector based on the saliency map for classification. Zhang *et al.* [14] propose to select the discriminative parts through exploiting the useful information in part clusters. In our OPAM approach, we first propose an

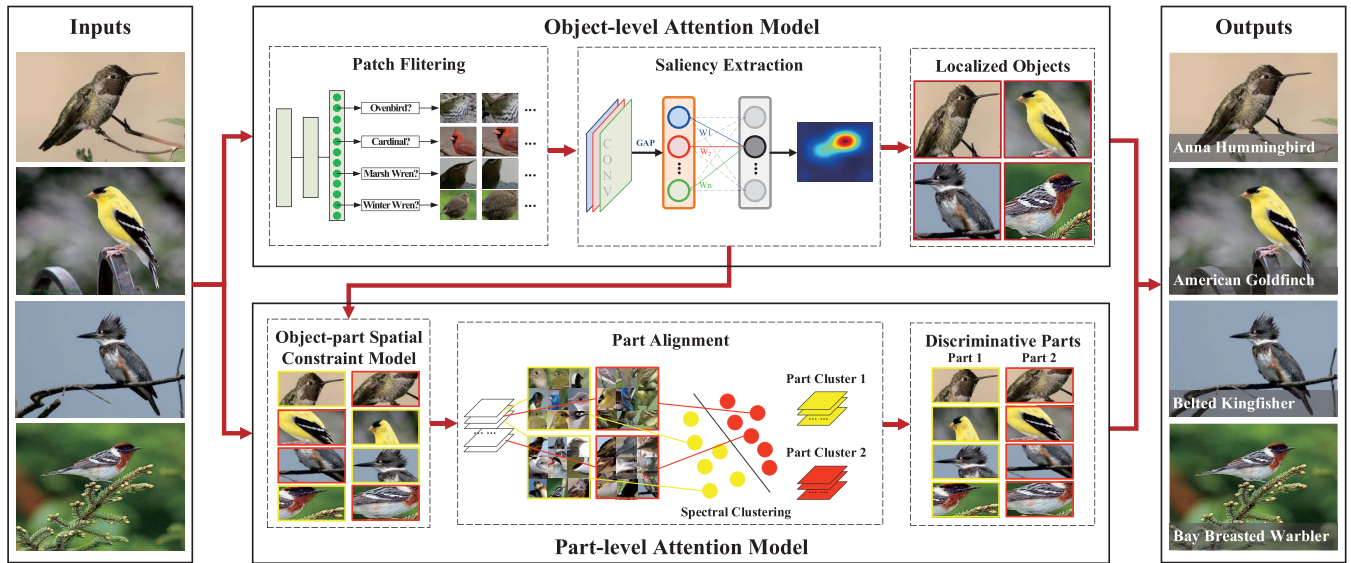


Fig. 3. An overview of our OPAM approach. The object-level attention model is to localize object for learning object features. The part-level attention model is to select the discriminative parts for exploiting the subtle and local features. The outputs show the predicted subcategories.

object-part spatial constraint model to select discriminative parts, which considers the spatial relationships between object and its parts as well as among these parts, and then utilizes the cluster patterns of neural network to align the parts with the same semantic meaning together for improving the classification performance.

III. OUR OPAM APPROACH

Our approach is based on an intuitive idea: fine-grained image classification generally first localizes the object (object-level attention) and then discriminative parts (part-level attention). For example, recognizing an image which contains a Field Sparrow follows the processes of first finding a bird, and then focusing on the discriminative parts that distinguish it from other bird subcategories. We propose the *object-part attention model* for weakly supervised fine-grained image classification, which uses neither object nor part annotations in both training and testing phases, and only uses the image-level subcategory labels. As shown in Fig. 3, our OPAM approach first localizes objects of images through object-level attention model for learning object features, and then selects the discriminative parts through part-level attention model for learning the subtle and local features. In the following subsections, the object-level and part-level attention models are presented respectively.

A. Object-Level Attention Model

Most existing weakly supervised works [7], [14], [20] devote to the discriminative part selection, but ignore the object localization, which can remove the influence of background noise in image to learn meaningful and representative object features. Although some methods consider both object localization and part selection, they rely on the object and part annotations [6], [19]. For addressing this important problem, we propose an object-level attention model based on the

saliency extraction for localizing the objects of images automatically only with image-level subcategory labels, without any object or part annotations. The model consists of two components: patch filtering and saliency extraction. The first component is to filter out the noisy image patches and retain those relevant to the object for training a CNN called *ClassNet*, to learn multi-view and multi-scale features for the specific subcategory. The second component is to extract the saliency map via global average pooling in CNN for localizing the objects of images.

1) *Patch Filtering*: A large amount of training data is significant for the performance of CNN, so we first focus on how to expand the training data. The bottom-up process can generate thousands of candidate image patches by grouping pixels into regions that may contain the object. These image patches can be used as the expansion of training data due to their relevances to the object. Therefore, selective search [9] is adopted to generate candidate image patches for a given image, which is an unsupervised and widely-used bottom-up process method. These candidate image patches provide multiple views and scales of original image, which benefit for training an effective CNN to achieve better fine-grained image classification accuracy. However, these patches can not be directly used due to the high recall but low precision, which means some noises exist. The object-level attention model is highly helpful for selecting the patches relevant to the object.

We remove the noisy patches and select relevant patches through a CNN, called *FilterNet*, which is pre-trained on the ImageNet 1K dataset [32], and then fine-tuned on the training data. We define the activation of neuron in softmax layer belonging to the subcategory of input image as the selection confidence score, and then a threshold is set to decide whether the given candidate image patch should be selected or not. Then we obtain the image patches relevant to the object with multiple views and scales. The expansion of training data improves the training effect of *ClassNet*, which has two

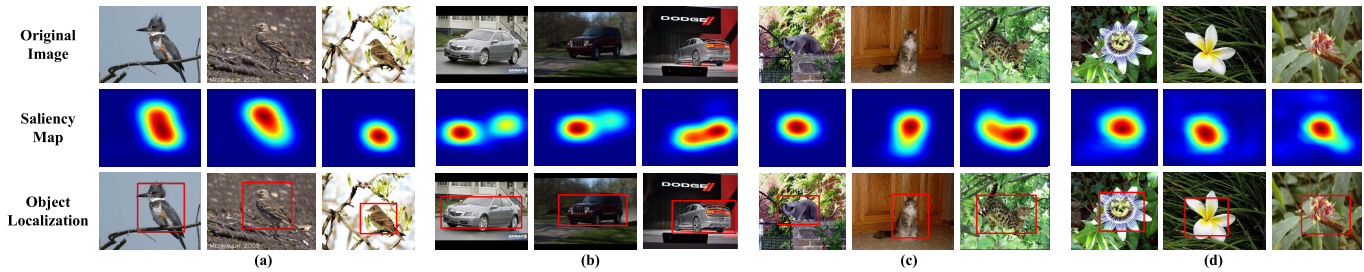


Fig. 4. Some results of saliency extraction by our OPAM approach. The first row shows the original images and the second row shows the saliency maps of original images. The object localization results are shown in the third row, in which the red rectangles represent the bounding boxes automatically produced by saliency extraction. The images in (a) Birds, (b) Cars, (c) Cats and (d) Flowers are from CUB-200-2011 [1], Cars-196 [2], Oxford-IIIT Pet [3] and Oxford-Flower-102 [4] datasets respectively.

aspects of benefits for our OPAM approach: (1) *ClassNet* is an effective fine-grained image classifier itself. (2) Its internal features are significantly helpful to build part clusters for aligning the parts with the same semantic meaning together, which will be described latter in Subsection B. It is noted that the patch filtering is performed only in the training phase and only uses image-level subcategory labels.

2) *Saliency Extraction*: In this stage, CAM [28] is adopted to obtain the saliency map M_c of an image for subcategory c to localize the object. The saliency map indicates the representative regions used by the CNN to identify the subcategory of image, as shown in the second row of Fig. 4. Then object regions of images, as shown in the third row of Fig. 4, are obtained by performing binarization and connectivity area extraction on the saliency maps.

Given an image I , the activation of neuron u in the last convolutional layer at spatial location (x, y) is defined as $f_u(x, y)$, and w_u^c defines the weight corresponding to subcategory c for neuron u . The saliency value at spatial location (x, y) for subcategory c is computed as follows:

$$M_c(x, y) = \sum_u w_u^c f_u(x, y) \quad (1)$$

where $M_c(x, y)$ directly indicates the importance of activation at spatial location (x, y) for classifying an image into subcategory c . Instead of using the image-level subcategory labels, we use the prediction result as the subcategory c in saliency extraction for each image. Through object-level attention model, we localize objects in the images to train a CNN called *ObjectNet* for obtaining the prediction of object-level attention.

B. Part-Level Attention Model

Since the discriminative parts, such as head and body, are crucial for fine-grained image classification, previous works [6] select discriminative parts from the candidate image patches produced by the bottom-up process like selective search [9]. However, these works rely on the part annotations which are heavily labor consuming. Although some works begin to focus on finding the discriminative parts without using any part annotations [7], [15], they ignore the spatial relationships between the object and its parts as well as among these parts. Therefore, we propose a new part selection

approach driven by part-level attention for exploiting the subtle and local discrimination to distinguish the subcategories, which uses neither object nor part annotations. It consists of two components: object-part spatial constraint model and part alignment. The first is to select the discriminative parts, and the second is to align the selected parts into clusters by the semantic meaning.

1) *Object-Part Spatial Constraint Model*: We obtain object regions of images through object-level attention model, and then employ object-part spatial constraint model to select the discriminative parts from the candidate image patches produced by the bottom-up process. Two spatial constraints are jointly considered: *object spatial constraint* defines the spatial relationship between object and its parts, and *part spatial constraint* defines the spatial relationship among these parts. For a given image I , its saliency map M and object region b are obtained through object-level attention model. Then part selection is driven by object-part spatial constraint model as follows:

Let \mathbb{P} denotes all the candidate image patches and $P = \{p_1, p_2, \dots, p_n\}$ denotes n parts selected from \mathbb{P} as the discriminative parts for each given image. The object-part spatial constraint model considers the combination of two spatial constraints by solving the following optimization problem:

$$P^* = \arg \max_{\mathbb{P}} \Delta(P) \quad (2)$$

where $\Delta(P)$ is defined as a scoring function over two spatial constraints as follows:

$$\Delta(P) = \Delta_{box}(P) \Delta_{part}(P) \quad (3)$$

Eq. 3 defines the proposed object-part spatial constraint, which ensures the representativeness and discrimination of the selected parts. It consists of two constraints: object spatial constraint $\Delta_{box}(P)$ and part spatial constraint $\Delta_{part}(P)$, which should be both satisfied by all the selected parts at the same time. For ensuring this, we choose product operation, not sum operation, as the work [6] which utilizes product operation to optimize two constraints.

a) *Object spatial constraint*: Ignoring the spatial relationship between the object and its parts causes that the selected parts may have large areas of background noise but small areas of discriminative region, which decreases the representativeness of selected parts. Since the discriminative

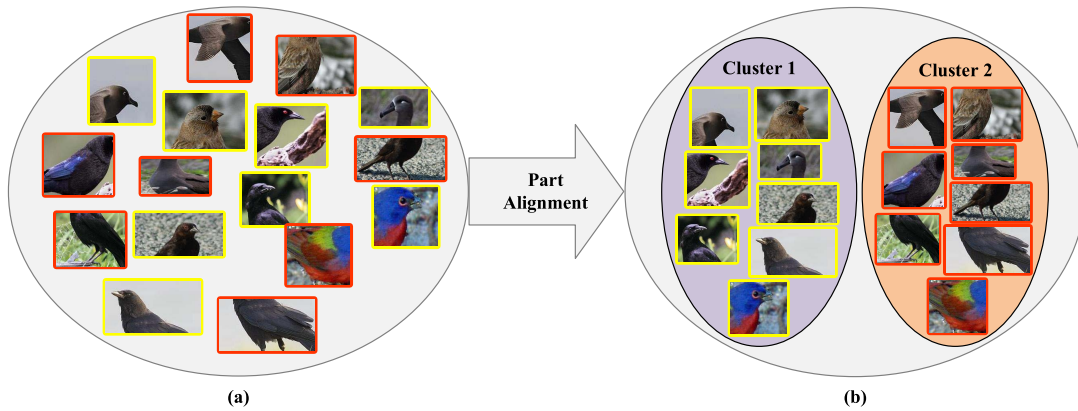


Fig. 5. Some results of part alignment in our OPAM approach. (a) shows the image patches which are selected through object-part spatial constraint model, and (b) shows that the image patches are aligned into clusters via part clusters.

parts are inside the object region, an intuitive spatial constraint function is defined as:

$$\Delta_{box}(P) = \prod_{i=1}^n f_b(p_i) \quad (4)$$

where

$$f_b(p_i) = \begin{cases} 1, & IoU(p_i) > threshold \\ 0, & otherwise \end{cases} \quad (5)$$

and $IoU(p_i)$ defines the proportion of Intersection-over-Union (IoU) overlap of part region and object region. It is noted that the object region is obtained automatically through the object-level attention model, *not provided by the object annotation*. Object spatial constraint aims to simultaneously restrain all the selected parts inside the object region. So product operation is utilized to ensure this, which is the same with the work [6]. That is to say, any part that does not satisfy object spatial constraint, e.g. its IoU value equals 0, will not be selected as a discriminative part.

b) Part spatial constraint: Ignoring the spatial relationship among these parts leads to the problem that the selected parts may have large overlap with each other, and some discriminative parts are ignored. The saliency map indicates the discrimination of image, and benefits for selecting discriminative parts. We jointly model saliency and the spatial relationship among parts as follows:

$$\Delta_{part}(P) = \log(A_U - A_I - A_O) + \log(\text{Mean}(M_{A_U})) \quad (6)$$

where A_U is the union area of n parts, A_I is the intersection area of n parts, A_O is the area outside the object region and $\text{Mean}(M_{A_U})$ is defined as follows:

$$\text{Mean}(M_{A_U}) = \frac{1}{|A_U|} \sum_{i,j} M_{ij} \quad (7)$$

where pixel (i, j) locates in the union area of parts, M_{ij} refers to the saliency value of pixel (i, j) , and $|A_U|$ refers to the number of pixels that locate in the union area of n parts. Part spatial constraint aims to select the most discriminative parts, which consists of two items: The first item aims to reduce the overlaps among selected parts, and is realized by

$\log(A_U - A_I - A_O)$, where $-A_I$ ensures the selected parts have the least overlap, and $-A_O$ ensures the selected parts have the largest areas inside the object region. The second item aims to maximize the saliency of selected parts, and is realized by $\log(\text{Mean}(M_{A_U}))$, which denotes the average saliency value of all the pixels in the union area of selected parts. We hope both of the two items in Eq. 6 have the maximum values, so sum operation is adopted.

2) Part Alignment: The selected parts through object-part spatial constraint model are in disorder and not aligned by its semantic meaning, as shown in Fig. 5(a). These parts with different semantic meanings contribute to the final prediction differently, so an intuitive idea is to align the parts with the same semantic meaning together, as shown in Fig. 5(b).

We are inspired by the fact that middle layers of *ClassNet* show clustering patterns. For example, there are groups of neurons significantly responding to the head of bird, and others to the body of bird, despite the fact that they may correspond to different poses. So clustering is performed on the neurons of a middle layer in the *ClassNet* to build the part clusters for aligning the selected parts. We first compute the similarity matrix S , where $S(i, j)$ denotes the cosine similarity of weights between two mid-layer neurons u_i and u_j , and then perform spectral clustering on the similarity matrix S to partition the mid-layer neurons into m groups. In the experiments, neurons are picked from the penultimate convolutional layer with m set as 2, as shown in Fig. 6, where the coordinate values represent the two largest eigenvectors of similarity matrices among all neurons, as the work [33].

Then we use the part clusters to align the selected parts as follows: (1) Warp the images of selected parts to the size of receptive field on input image of neuron in penultimate convolutional layer. (2) Feed forward the selected parts to the penultimate convolutional layer to produce an activation score for each neuron. (3) Sum up the scores of neurons in one cluster to get cluster score. (4) Align the selected parts to the cluster with highest cluster score, which is formulated as follows: For a given image, n discriminative parts $P = \{p_1, p_2, \dots, p_n\}$ are obtained by object-part spatial constraint model, and then part alignment is performed on these parts

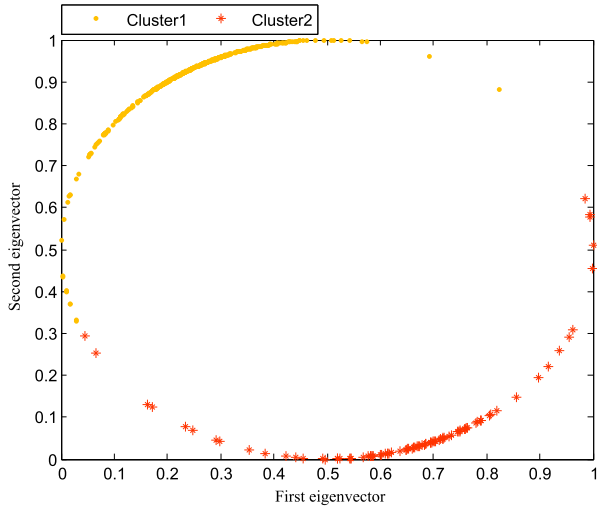


Fig. 6. Illustration of spectral clustering. The coordinate values represent the two largest eigenvectors of similarity matrices among all neurons.

Algorithm 1 Part Alignment

Input: The i th selected part p_i ; The part clusters $L = \{l_1, l_2, \dots, l_m\}$; And the number of neurons in penultimate convolutional layer d .

Output: The cluster that p_i is aligned into l_c .

- 1: Set $score_k = 0; k = 1, \dots, m$.
 - 2: Warp p_i to the size of receptive field on input image of neuron in penultimate convolutional layer.
 - 3: Perform a feed-forward pass to compute p_i 's activations $F_i = \{f_{i1}, f_{i2}, \dots, f_{id}\}$.
 - 4: **for** $k = 1, \dots, m; j = 1, \dots, d$ **do**
 - 5: **if** j th neuron belongs to cluster l_k **then**
 - 6: $score_k = score_k + f_{ij}$.
 - 7: **end if**
 - 8: **end for**
 - 9: $c = \arg \max_k score_k$.
 - 10: **return** l_c .
-

with m part clusters $L = \{l_1, l_2, \dots, l_m\}$ as Algorithm 1.

The choice of middle layer has important influence on the part alignment and classification performance. We follow standard practice and withhold a validation set of 10% training data for grid search to determine which layer to choose. At last, we find the penultimate convolutional layer works better than others. Through part-level attention model, we select the discriminative parts in images to train a CNN called *PartNet* for obtaining the prediction of part-level attention.

C. Final Prediction

For better classification performance, we fine-tune *ClassNet* with the localized object and the discriminative parts to get two classifiers, called *ObjectNet* and *PartNet* respectively. *ClassNet*, *ObjectNet* and *PartNet* are all fine-grained image classifiers: *ClassNet* for original images, *ObjectNet* for objects and *PartNet* for selected discriminative parts. However, their impacts and strengths are different, primarily because they focus on the different natures of image.

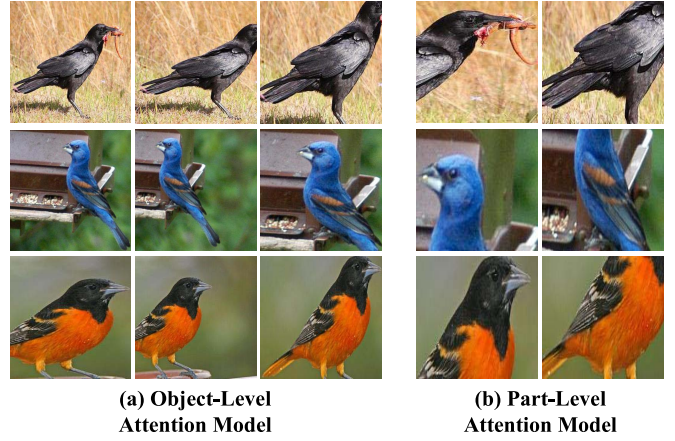


Fig. 7. Some results of selected image patches by the object-level and part-level attention model respectively. Image patches selected by the object-level attention model focus on the whole objects, as shown in (a). Image patches selected by the part-level attention model focus on subtle and local features, as shown in (b).

Object-level attention model first drives *FilterNet* to select image patches with multiple views and scales that are relevant to the object, as shown in Fig. 7 (a). These image patches drive *ClassNet* to learn more representative features and localize the object region through saliency extraction. Part-level attention model selects discriminative parts which contain subtle and local features, as shown in Fig. 7 (b). The different level focuses (i.e. original image, object of original image, and parts of original image) have different representations and are complementary to improve the prediction. Finally, we merge the prediction results of the three different levels by using the following equation:

$$final_score = \alpha * original_score + \beta * object_score + \gamma * part_score \quad (8)$$

where $original_score$, $object_score$ and $part_score$ are the softmax values of *ClassNet*, *ObjectNet* and *PartsNet* respectively, and α , β and γ are selected by using the k -fold cross-validation method [34]. The subcategory with the highest $final_score$ is chosen as the final prediction result.

IV. EXPERIMENTS

We conduct experiments on four widely-used datasets for fine-grained image classification: CUB-200-2011, Cars-196, Oxford-IIIT Pet and Oxford-Flower-102. Our proposed OPAM approach is compared with more than ten state-of-the-art methods to verify its effectiveness.

A. Datasets and evaluation metric

Four datasets are adopted for the experiments:

- **CUB-200-2011** [1]: It is the most widely-used dataset for fine-grained image classification, and contains 11788 images of 200 different bird subcategories, which is divided as follows: 5994 images for training and 5794 images for testing. For each subcategory, 30 images are selected for training and 11~30 images for testing,

and each image has detailed annotations: a subcategory label, a bounding box of object, 15 part locations and 312 binary attributes. All attributes are visual in nature, pertaining to color, pattern, or shape of a particular part.

- **Cars-196** [2]: It contains 16185 images of 196 car subcategories, and is divided as follows: 8144 images for training and 8041 images for testing. For each subcategory, 24~84 images are selected for training and 24~83 images for testing. Each image is annotated with a subcategory label and a bounding box of object.
- **Oxford-IIT Pet** [3]: It is a collection of 7349 images with 37 different pet subcategories, among which 12 are cat subcategories and 25 are dog subcategories. It is divided as follows: 3680 images for training and 3669 images for testing. For each subcategory, 93~100 images are selected for training and 88~100 images for testing. Each image is annotated with a subcategory label, a pixel level segmentation marking the body and a tight bounding box of head.
- **Oxford-Flower-102** [4]: It has 8189 images of 102 subcategories belonging to flowers, 1020 for training, 1020 for validation and 6149 for testing. One image may contain several flowers. Each image is annotated with a subcategory label.

Accuracy is adopted as the evaluation metric to comprehensively evaluate the classification performances of our OPAM approach and compared methods, which is widely used for evaluating the performance of fine-grained image classification [6], [7], [14], and is defined as follows:

$$Accuracy = \frac{R_a}{R} \quad (9)$$

where R means the number of testing images and R_a counts the number of images which are correctly classified.

B. Details of the networks

In the experiments, the widely-used CNN of VGGNet [34] is adopted. It is noted that the CNN used in our proposed approach can be replaced with the other CNNs. In our approach, CNN has two different effects: localization and classification. Therefore, the architectures of CNNs are modified for different functions:

1) *Localization*: In the object-level attention model, CNN is used to extract the saliency map of an image for object localization. Zhou *et al.* [28] find that the accuracy of localization can be improved if the last convolutional layer before global average pooling has a higher spatial resolution, which is termed as the mapping resolution. In order to get a higher spatial resolution, the layers after conv5_3 are removed, resulting in a mapping resolution of 14×14 . Besides, a convolutional layer of size 3×3 , stride 1, pad 1 with 1024 neurons is added, followed by a global average pooling layer and a softmax layer. The modified VGGNet is pre-trained on the 1.3M training images of ImageNet 1K dataset [32], and then fine-tuned on the fine-grained image classification dataset. The number of neurons in softmax layer is set as the number of subcategories.

2) *Classification*: The CNN used in the experiments for classification is the VGGNet [34] with batch normalization [35]. For the prediction results of original image, object and parts, the same CNN architecture is used but fine-tuned on different training data. For the prediction of original image, we fine-tune the CNN on the image patches selected through object-level attention model, as *ClassNet*. For the predictions of object and part, we fine-tuned the CNNs on the images of objects and images of parts based on *ClassNet* respectively, as *ObjectNet* and *PartNet*. Then we can get prediction results of the three different levels in Eq. 8. We follow the work [6] to select the 3 parameters (i.e. α , β and γ) by k -fold cross-validation method [34]. Considering that the scale of training dataset is small, we set k as 3 to ensure that each subset of the training dataset is not too small, which guarantees a better selection of parameters. We follow [34] to randomly split the training dataset D into 3 mutually exclusive subsets D_1, D_2, D_3 of equal size. We conduct experiment 3 times. For each time t , we train on $D \setminus D_t$ and test on D_t . For parameter selection, we traverse the value of each parameter from 0 to 1 by step of 0.1. We select the parameters that obtain the highest classification accuracy. Finally, for CUB-200-2011, Cars-196, Oxford-IIT Pet and Oxford-Flower-102 datasets, (α, β, γ) are set as (0.4, 0.4, 0.2), (0.5, 0.3, 0.2), (0.4, 0.4, 0.2) and (0.4, 0.3, 0.3).

C. Comparisons with the state-of-the-art methods

This subsection presents the experimental results and analyses of our OPAM approach on four widely-used fine-grained image classification datasets as well as the state-of-the-art methods. Table I shows the comparison results on CUB-200-2011 dataset. The object, part annotations and CNN features used in these methods are listed for fair comparison. CNN models shown in the column of ‘‘CNN Features’’, such as VGGNet [34] and GoogleNet [61], indicate which CNN model this method adopts to extract CNN features. If the column is empty, it means that the result of this method is produced by handcrafted feature like SIFT.

Early works [38], [43], [44] choose SIFT [16] as features, and the performances are limited and much lower than our OPAM approach no matter whether using the object and part annotations or not. Our approach is the best among all methods under the same setting that neither object nor part annotations are used in both training and testing phases, and obtains 1.20% higher accuracy than the best compared result of FOAF [8] (85.83% vs. 84.63%). It is noted that the CNN used in FOAF is pre-trained not only on ImageNet 1K dataset [32] but also on the dataset of PASCAL VOC [60], while our approach does not use the external dataset like PASCAL VOC. Compared with the second highest result of PD [7], our approach achieves 1.29% higher accuracy (85.83% vs. 84.54%). Our OPAM approach improves 7.93% than our previous conference paper [15], and it verifies the effectiveness of further exploitation in our OPAM approach, which jointly integrates the object-level and part-level attention models to boost the multi-view and multi-scale feature learning and enhance their complementarity. Besides, our OPAM

TABLE I
COMPARISONS WITH STATE-OF-THE-ART METHODS ON CUB-200-2011 DATASET

Method	Train Annotation		Test Annotation		Accuracy (%)	CNN Features
	Object	Parts	Object	Parts		
Our OPAM Approach					85.83	VGGNet
FOAF [8]					84.63	VGGNet
PD [7]					84.54	VGGNet
STN [21]					84.10	GoogleNet
Bilinear-CNN [25]					84.10	VGGNet&VGG-M
Multi-grained [24]					81.70	VGGNet
NAC [20]					81.01	VGGNet
PIR [14]					79.34	VGGNet
TL Atten [15]					77.90	VGGNet
MIL [37]					77.40	VGGNet
VGG-BGLm [13]					75.90	VGGNet
InterActive [30]					75.62	VGGNet
Dense Graph Mining [38]					60.19	
Coarse-to-Fine [39]	✓				82.50	VGGNet
Coarse-to-Fine [39]	✓		✓		82.90	VGGNet
PG Alignment [12]	✓		✓		82.80	VGGNet
VGG-BGLm [13]	✓		✓		80.40	VGGNet
Triplet-A (64) [40]	✓		✓		80.70	GoogleNet
Triplet-M (64) [40]	✓		✓		79.30	GoogleNet
Webly-supervised [41]	✓	✓			78.60	AlexNet
PN-CNN [11]	✓	✓			75.70	AlexNet
Part-based R-CNN [6]	✓	✓			73.50	AlexNet
SPDA-CNN [23]	✓	✓	✓		85.14	VGGNet
Deep LAC [42]	✓	✓	✓		84.10	AlexNet
SPDA-CNN [23]	✓	✓	✓		81.01	AlexNet
PS-CNN [22]	✓	✓	✓		76.20	AlexNet
PN-CNN [11]	✓	✓	✓	✓	85.40	AlexNet
Part-based R-CNN [6]	✓	✓	✓	✓	76.37	AlexNet
POOF [43]	✓	✓	✓	✓	73.30	
HPM [44]	✓	✓	✓	✓	66.35	

TABLE II
COMPARISONS WITH STATE-OF-THE-ART METHODS ON CARS-196 DATASET

Method	Train Annotation		Test Annotation		Accuracy (%)	CNN Features
	Object	Parts	Object	Parts		
Our OPAM Approach					92.19	VGGNet
Bilinear-CNN [25]					91.30	VGGNet&VGG-M
TL Atten [15]					88.63	VGGNet
DVAN [31]					87.10	VGGNet
FT-HAR-CNN [45]					86.30	AlexNet
HAR-CNN [45]					80.80	AlexNet
PG Alignment [12]	✓				92.60	VGGNet
ELLF [46]	✓				73.90	CNN
R-CNN [10]	✓				57.40	AlexNet
PG Alignment [12]	✓		✓		92.80	VGGNet
BoT(CNN With Geo) [47]	✓		✓		92.50	VGGNet
DPL-CNN [48]	✓		✓		92.30	VGGNet
VGG-BGLm [13]	✓		✓		90.50	VGGNet
LLC [49]	✓		✓		69.50	
BB-3D-G [2]	✓		✓		67.60	

approach employs the object-part spatial constraint model to exploit the subtle and local discrimination for distinguishing the subcategories.

Our approach performs better than the methods which focus on the CNN architectures, such as STN [21] and Bilinear-CNN [25]. In STN, GoogleNet [61] with batch normalization [35] is adopted to achieve the accuracy of 82.30% by only fine-tuning on CUB-200-2011 dataset without any other processing. Two different CNNs are employed in Bilinear-CNN: VGGNet [34] and VGG-M [62]. The classification accuracies of the two methods are both 84.10%, which are lower than our approach by 1.73%.

Furthermore, our approach outperforms the methods which use object annotation, such as Coarse-to-Fine [39], PG Alignment [12] and VGG-BGLm [13]. Moreover, our approach outperforms methods that use both object and part annotations [6], [23]. Neither object nor part annotations are used in our OPAM approach, which makes fine-grained image classification march toward practical application.

Besides, the results on Cars-196, Oxford-IIIT Pet and Oxford-Flower-102 datasets are shown in Tables II, III and IV respectively. The trends of results on these three datasets are similar as CUB-200-2011 dataset, our OPAM approach achieves the best results among state-of-the-art methods

TABLE III
COMPARISONS WITH STATE-OF-THE-ART METHODS ON OXFORD-IIIT PET DATASET

Method	Accuracy (%)	CNN Features
Our OPAM Approach	93.81	VGGNet
InterActive [30]	93.45	VGGNet
TL Atten [15]	92.51	VGGNet
NAC [20]	91.60	VGGNet
FOAF [8]	91.39	VGGNet
ONE+SVM [50]	90.03	VGGNet
Deep Optimized [51]	88.10	AlexNet
NAC [20]	85.20	AlexNet
MsML+ [52]	81.18	CNN
MsML [52]	80.45	CNN
Deep Standard [51]	78.50	AlexNet
Shape+Appearance [3]	56.68	
Zernike+SCC [53]	59.50	
GMP+p [54]	56.80	
GMP [54]	56.10	
M-HMP [55]	53.40	
Detection+Segmentation [56]	54.30	

TABLE IV
COMPARISONS WITH STATE-OF-THE-ART METHODS ON OXFORD-FLOWER-102 DATASET

Method	Accuracy (%)	CNN Features
Our OPAM Approach	97.10	VGGNet
InterActive [30]	96.40	VGGNet
PBC [57]	96.10	GoogleNet
TL Atten [15]	95.76	VGGNet
NAC [20]	95.34	VGGNet
RIIR [58]	94.01	VGGNet
Deep Optimized [51]	91.30	AlexNet
Deep Standard [51]	90.50	AlexNet
MML [52]	89.45	CNN
CNN Feature [59]	86.80	CNN
Generalized Max Pooling [54]	84.60	
Detection+Segmentation [56]	80.66	

TABLE V
PERFORMANCES OF COMPONENTS IN OUR OPAM APPROACH ON CUB-200-2011, CARS-196, OXFORD-IIIT PET AND OXFORD-FLOWER-102 DATASETS

Method	Accuracy (%)			
	CUB-200-2011	Cars-196	Oxford-IIIT Pet	Oxford-Flower-102
Our OPAM Approach (Original+Object-level+Part-level)	85.83	92.19	93.81	97.10
Original	80.82	86.79	88.14	94.70
Object-level	83.74	88.79	90.98	95.32
Part-level	80.65	84.26	85.75	93.09
Original+Object-level	84.79	91.15	92.20	96.55
Original+Part-level	84.41	91.06	91.82	96.23
Object-level+Part-level	84.73	89.69	91.50	95.66

(92.19%, 93.81% and 97.10% respectively) and brings 0.89%, 0.36% and 0.70% improvements than the best results of compared methods respectively.

D. Performances of components in our OPAM approach

Detailed experiments are performed on our OPAM approach from the following three aspects:

1) *Effectivenesses of object-level attention and part-level attention models*: In our OPAM approach, the final prediction score is generated by merging the prediction scores of three different images, i.e. original image, image of object and images of parts, which are denoted as “Original”, “Object-level” and “Part-level”. The effectivenesses of object-level

and part-level attention models are verified in the following paragraphs. From Table V, Fig. 8 and 9, we can observe that:

- Object-level attention model improves the classification accuracy via localizing objects of images for learning global features. Comparing with the result of “Original”, it improves by 2.92%, 2.00%, 2.84% and 0.62% on four datasets respectively, and combining “Object-level” with “Original” improves even more, i.e. by 3.97%, 4.36%, 4.06% and 1.85% on four datasets respectively. The classification accuracy of part-level attention model is not higher than “Original”. Fig. 9 shows some failure results of part selection. We conclude that our proposed part selection approach may fail in following two cases: 1) Objects are hard to be distinguished from the back-

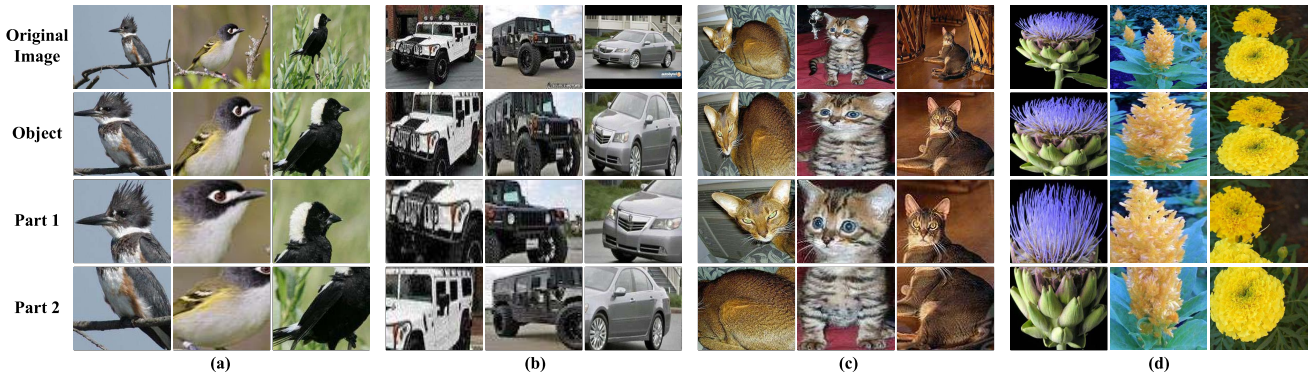


Fig. 8. Some results of object localization and part selection. The first row denotes the original images, the second row denotes the localized objects of original images via object-level attention model, and the third and fourth rows denote the selected discriminative parts via part-level attention model. The images in (a) Birds, (b) Cars, (c) Cats and (d) Flowers are from CUB-200-2011 [1], Cars-196 [2], Oxford-IIIT Pet [3] and Oxford-Flower-102 [4] datasets respectively.

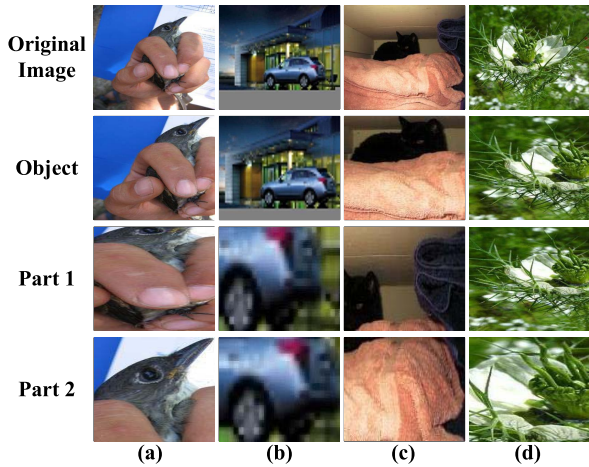


Fig. 9. Some failure results of part selection. The images in (a) Birds, (b) Cars, (c) Cats and (d) Flowers are from CUB-200-2011 [1], Cars-196 [2], Oxford-IIIT Pet [3] and Oxford-Flower-102 [4] datasets respectively.

ground. 2) Objects are in heavy occlusion. In these two cases, it is hard to localize the object accurately so the part selection fails, which is based on the object localization. The failure of part selection is the first reason of lower accuracy only with part. Another reason is that part-level attention focuses on the subtle and local features of object, containing less information than original image. However, despite these challenging cases, “Part-level” still achieves considerable classification accuracies, which is better than some state-of-the-art methods, such as [13] and [37]. Besides, it is complementary with original image and object, so their combination further boosts the classification accuracy and achieves the best result compared with state-of-the-art methods.

- Combining object-level and part-level attention models achieves more accurate results than only one level attention model, e.g. 84.73% vs. 83.74% and 80.65% on CUB-200-2011 dataset. Combining the two level attention models with “Original” improves a lot than “Original”, i.e. by 5.01%, 5.40%, 5.67% and 2.4% on the

four datasets respectively. It shows the complementarity of object-level and part-level attention models in fine-grained image classification. The two level attention models have different but complementary focuses: the object-level attention model focuses on differences of representative object appearances, while the part-level attention model focuses on the subtle and local differences of discriminative parts among subcategories. Both of them are jointly employed to boost the multi-view and multi-scale feature learning and enhance their mutual promotion to achieve better performance for fine-grained image classification.

- We observe that “Original+Part-level” is better than “Object-level+Part-level”, which shows the complementarity between “Original” and “Part-level”. This is because: 1) Parts are selected based on the obtained object regions, which leads to the fact that selected parts are mostly inside object regions and cover the whole object regions. This causes that the complementarity between object and part is small. 2) Object localization may be wrong and cause that the localized object region does not contain the whole object, some areas of this object are outside the object region. These areas may be helpful for classification, which are not in the localized object region but in the original image. 3) Image also includes the information of background, which may be helpful for classification to a certain extent. So “Original+Part-level” can provide more supplementary information than “Object-level+Part-level”, thus achieves better performance. Totally, “Original+Object-level+Part-level” further improves the classification accuracy due to the complementary information among image, object and part.
- Fig. 8 shows some results of object localization and part selection by our OPAM approach. The first row denotes the original images, the second row denotes the localized objects of original images via object-level attention model, and the third and fourth rows denote the selected discriminative parts via part-level attention

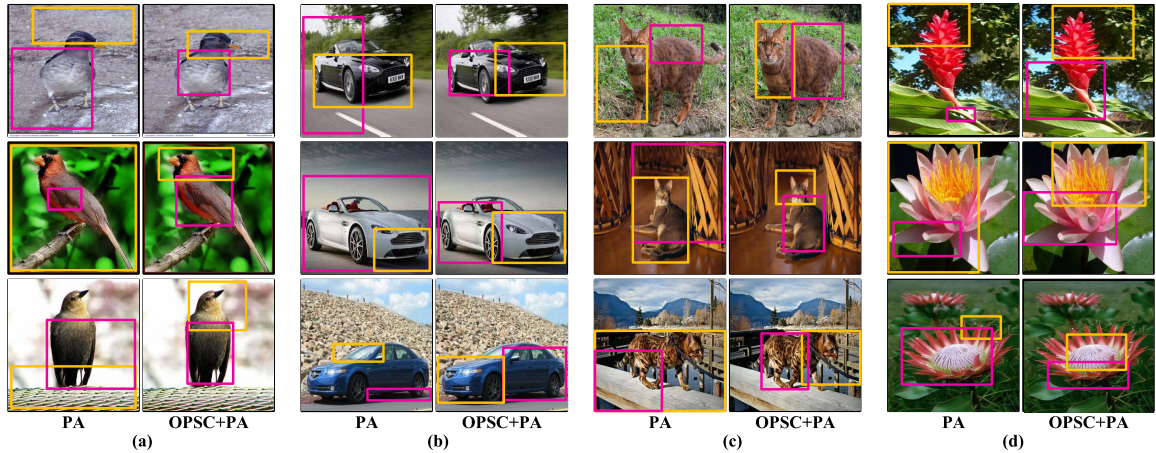


Fig. 10. Examples of part selection from our previous conference paper [15] (left column) and our OPAM approach in this paper (right column). “PA” refers to part alignment which is adopted in our previous conference paper [15], “OPSC” refers to object-part spatial constraint model, and “OPSC+PA” refers to combining the above two approaches, which is adopted in our OPAM approach. The yellow and orange rectangles denote the selected discriminative parts via the two approaches, which respond to the heads and bodies of objects. The images in (a) Birds, (b) Cars, (c) Cats and (d) Flowers are from CUB-200-2011 [1], Cars-196 [2], Oxford-IIIT Pet [3] and Oxford-Flower-102 [4] datasets respectively.

TABLE VI

PERFORMANCES OF OBJECT-PART SPATIAL CONSTRAINT MODEL, PART ALIGNMENT AND THEIR COMBINATION

Method	Accuracy (%)			
	CUB-200-2011	Cars-196	Oxford-IIIT Pet	Oxford-Flower-102
OPSC+PA (ours)	80.65	84.26	85.75	93.09
OPSC (ours)	79.74	83.34	83.46	92.33
PA (our previous [15])	65.41	68.32	75.42	88.75

model. For CUB-200-2011, Cars-196 and Oxford-IIIT Pet datasets, the selected parts have explicit semantic meanings, where the third row denotes the head of object and the fourth denotes the body. For Oxford-Flower-102 dataset, there are two types of images: one contains only one flower, and the other contains multiple flowers. For the images containing only one flower, object means the flower and parts mean the discriminative regions of the flower, such as petal, flower bud or receptacle. For the images containing multiple flowers, object means the salient flower or the entirety of all flowers in image, and parts mean the discriminative regions of the flower or one single individual of the flowers. Our proposed approach is effective in both two cases, which localizes the discriminative objects and parts as well as learns fine-grained features to boost classification accuracy. It is noted that neither object nor part annotations are used in our OPAM approach, which avoids the heavy labor consumption of labeling as well as pushes fine-grained image classification towards practical applications.

2) *Effectiveness of object-part spatial constraint model and part alignment*: Compared with our previous conference paper [15], which only performs part alignment for selecting discriminative parts, we further employ object-part spatial constraint model to drive the discriminative part selection. The object spatial constraint ensures selected parts with high representativeness, while part spatial constraint eliminates redundancy and enhances discrimination of selected parts. Both of them are jointly employed to exploit the subtle and local discrimination for distinguishing the subcategories. In Fig. 10 and

Table VI, “OPSC” refers to the object-part spatial constraint model, “PA” refers to part alignment which is adopted by our previous conference paper [15], and “OPSC+PA” refers to combining the above two ones, which is adopted by our OPAM approach. From the left columns of four datasets in Fig. 10, we can see that only performing part alignment in part-level attention model without object-part spatial constraint causes the selected parts: (1) have large areas of background noise but small areas of object, (2) have large overlap with each other which leads to the redundant information. From Table VI, we can see that the classification accuracies of parts selected by object-part spatial constraint model (“OPSC”) are better than parts selected with part alignment (“PA”) on all 4 datasets. Besides, applying part alignment on the basis of object-part spatial constraint further improves the classification performance. This verifies that aligning discriminative parts with the same semantic meaning together can further improve the results of part-level attention model.

3) *Effectiveness of patch filtering*: Through patch filtering in the object-level attention model, some image patches are selected from the candidate image patches. These patches are relevant to objects, and provide multiple views and scales of original images. These relevant patches are used to train *ClassNet* to boost the effectiveness of *ClassNet*. In Table VII, “ft-patches” refers to fine-tuning on image patches selected through patch filtering in object-level attention model and “ft-original” refers to fine-tuning only on original images. The results are the classification accuracies of prediction on original images. Fine-tuning on the selected image patches achieves better accuracy due to the effectiveness of multi-view

TABLE VII
PERFORMANCES OF PATCH FILTERING

Method	Accuracy (%)			
	CUB-200-2011	Cars-196	Oxford-IIIT Pet	Oxford-Flower-102
ft-patches	80.82	86.79	88.14	94.70
ft-original	80.11	85.76	87.52	93.84

and multi-scale feature learning based on the patch filtering in our OPAM approach.

V. CONCLUSION

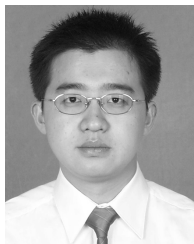
In this paper, the OPAM approach has been proposed for weakly supervised fine-grained image classification, which jointly integrates two level attention models: object-level localizes objects of images, and part-level selects discriminative parts of objects. The two level attentions jointly improve the multi-view and multi-scale feature learning and enhance their mutual promotion. Besides, part selection is driven by the object-part spatial constraint model, which combines two spatial constraints: object spatial constraint ensures the high representativeness of selected parts, and part spatial constraint eliminates redundancy and enhances discrimination of selected parts. Combination of the two spatial constraints promotes the subtle and local discrimination localization. Importantly, our OPAM avoids the heavy labor consumption of labeling to march toward practical application. Comprehensive experimental results show the effectiveness of our OPAM approach compared with more than ten state-of-the-art methods on four widely-used datasets.

The future work lies in two aspects: First, we will focus on learning better fine-grained representation via more effective and precise part localization methods. Second, we will also attempt to apply semi-supervised learning into our work to make full use of large amounts of web data. Both of them will be employed to further improve the fine-grained image classification performance.

REFERENCES

- [1] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-UCSD birds-200-2011 dataset," California Inst. Technol., Pasadena, CA, USA, Tech. Rep. CNS-TR-2011-001, 2011.
- [2] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3D object representations for fine-grained categorization," in *Proc. Int. Conf. Comput. Vis. Workshop (ICCV)*, 2013, pp. 554–561.
- [3] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar, "Cats and dogs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 3498–3505.
- [4] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proc. 6th Indian Conf. Comput. Vis., Graph. Image Process.*, Dec. 2008, pp. 722–729.
- [5] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi. (Jun. 2013). "Fine-grained visual classification of aircraft." [Online]. Available: <https://arxiv.org/abs/1306.5151>
- [6] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based R-CNNs for fine-grained category detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 834–849.
- [7] X. Zhang, H. Xiong, W. Zhou, W. Lin, and Q. Tian, "Picking deep filter responses for fine-grained image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1134–1142.
- [8] X. Zhang, H. Xiong, W. Zhou, and Q. Tian, "Fused one-vs-all features with semantic alignments for fine-grained visual categorization," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 878–892, Feb. 2016.
- [9] J. R. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, 2013.
- [10] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 580–587.
- [11] S. Branson, G. Van Horn, S. Belongie, and P. Perona. (Jun. 2014). "Bird species categorization using pose normalized deep convolutional nets." [Online]. Available: <https://arxiv.org/abs/1406.2952>.
- [12] J. Krause, H. Jin, J. Yang, and L. Fei-Fei, "Fine-grained recognition without part annotations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5546–5555.
- [13] F. Zhou and Y. Lin, "Fine-grained image classification by exploring bipartite-graph labels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1124–1133.
- [14] Y. Zhang *et al.*, "Weakly supervised fine-grained categorization with part-based image representation," *IEEE Trans. Image Process.*, vol. 25, no. 4, pp. 1713–1725, Apr. 2016.
- [15] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang, "The application of two-level attention models in deep convolutional neural network for fine-grained image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 842–850.
- [16] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
- [17] L. Xie, Q. Tian, M. Wang, and B. Zhang, "Spatial pooling of heterogeneous features for image classification," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 1994–2008, May 2013.
- [18] S. Gao, I. W.-H. Tsang, and Y. Ma, "Learning category-specific dictionary and shared dictionary for fine-grained image categorization," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 623–634, Feb. 2014.
- [19] N. Zhang, R. Farrell, F. Iandola, and T. Darrell, "Deformable part descriptors for fine-grained recognition and attribute prediction," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, 2013, pp. 729–736.
- [20] M. Simon and E. Rodner, "Neural activation constellations: Unsupervised part model discovery with convolutional networks," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 1143–1151.
- [21] M. Jaderberg *et al.*, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 2017–2025.
- [22] S. Huang, Z. Xu, D. Tao, and Y. Zhang, "Part-stacked CNN for fine-grained visual categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1173–1182.
- [23] H. Zhang *et al.*, "SPDA-CNN: Unifying semantic part detection and abstraction for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1143–1152.
- [24] D. Wang, Z. Shen, J. Shao, W. Zhang, X. Xue, and Z. Zhang, "Multiple granularity descriptors for fine-grained categorization," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 2399–2406.
- [25] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear CNN models for fine-grained visual recognition," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 1449–1457.
- [26] B. Zhao, J. Feng, X. Wu, and S. Yan, "A survey on deep learning-based fine-grained object classification and semantic segmentation," *Int. J. Autom. Comput.*, vol. 14, no. 2, pp. 119–135, 2017.
- [27] P. Sermanet, A. Frome, and E. Real. (Dec. 2014). "Attention for fine-grained categorization." [Online]. Available: <https://arxiv.org/abs/1412.7054>
- [28] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.
- [29] X. Liu, T. Xia, J. Wang, and Y. Lin. (Mar. 2016). "Fully convolutional attention networks for fine-grained recognition." [Online]. Available: <https://arxiv.org/abs/1603.06765>
- [30] L. Xie, L. Zheng, J. Wang, A. L. Yuille, and Q. Tian, "InterActive: Inter-layer activeness propagation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 270–279.
- [31] B. Zhao, X. Wu, J. Feng, Q. Peng, and S. Yan, "Diversified visual attention networks for fine-grained object classification," *IEEE Trans. Multimedia*, vol. 19, no. 6, pp. 1245–1256, Jun. 2017.

- [32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 248–255.
- [33] B. Nadler, S. Lafon, I. G. Kevrekidis, and R. R. Coifman, "Diffusion maps, spectral clustering and eigenfunctions of Fokker-Planck operators," in *Proc. 18th Int. Conf. Adv. Neural Inf. Process. Syst. (NIPS)*, 2006, pp. 955–962.
- [34] K. Simonyan and A. Zisserman. (Sep. 2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: <https://arxiv.org/abs/1409.1556>
- [35] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2015, pp. 448–456.
- [36] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 1995, vol. 14, no. 2, pp. 1137–1145.
- [37] Z. Xu, D. Tao, S. Huang, and Y. Zhang, "Friend or foe: Fine-grained categorization with weak supervision," *IEEE Trans. Image Process.*, vol. 26, no. 1, pp. 135–146, Jan. 2017.
- [38] L. Zhang, Y. Yang, M. Wang, R. Hong, L. Nie, and X. Li, "Detecting densely distributed graph patterns for fine-grained image categorization," *IEEE Trans. Image Process.*, vol. 25, no. 2, pp. 553–565, Feb. 2016.
- [39] H. Yao, S. Zhang, Y. Zhang, J. Li, and Q. Tian, "Coarse-to-fine description for fine-grained visual categorization," *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4858–4872, Oct. 2016.
- [40] Y. Cui, F. Zhou, Y. Lin, and S. Belongie, "Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1153–1162.
- [41] Z. Xu, S. Huang, Y. Zhang, and D. Tao, "Webly-supervised fine-grained visual categorization via deep domain adaptation," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: [10.1109/TPAMI.2016.2637331](https://doi.org/10.1109/TPAMI.2016.2637331).
- [42] D. Lin, X. Shen, C. Lu, and J. Jia, "Deep LAC: Deep localization, alignment and classification for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1666–1674.
- [43] T. Berg and P. N. Belhumeur, "POOF: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 955–962.
- [44] L. Xie, Q. Tian, R. Hong, S. Yan, and B. Zhang, "Hierarchical part matching for fine-grained visual categorization," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, 2013, pp. 1641–1648.
- [45] S. Xie, T. Yang, X. Wang, and Y. Lin, "Hyper-class augmented and regularized deep learning for fine-grained image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2645–2654.
- [46] J. Krause, T. Gebu, J. Deng, L.-J. Li, and L. Fei-Fei, "Learning features and parts for fine-grained recognition," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, 2014, pp. 26–33.
- [47] Y. Wang, J. Choi, V. Morariu, and L. S. Davis, "Mining discriminative triplets of patches for fine-grained classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1163–1172.
- [48] Y. Wang, V. I. Morariu, and L. S. Davis. (Nov. 2016). "Weakly-supervised discriminative patch learning via CNN for fine-grained recognition." [Online]. Available: <https://arxiv.org/abs/1611.09932>
- [49] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 3360–3367.
- [50] L. Xie, R. Hong, B. Zhang, and Q. Tian, "Image classification and retrieval are ONE," in *Proc. ACM Int. Conf. Multimedia Retr. (ICMR)*, 2015, pp. 3–10.
- [51] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "From generic to specific deep representations for visual recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPR)*, Jun. 2015, pp. 36–45.
- [52] Q. Qian, R. Jin, S. Zhu, and Y. Lin, "Fine-grained visual categorization via multi-stage metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3716–3724.
- [53] A. Iscen, G. Tolias, P.-H. Gosselin, and H. Jégou, "A comparison of dense region detectors for image search and fine-grained classification," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2369–2381, Aug. 2015.
- [54] N. Murray and F. Perronnin, "Generalized max pooling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 2473–2480.
- [55] L. Bo, X. Ren, and D. Fox, "Multipath sparse coding using hierarchical matching pursuit," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 660–667.
- [56] A. Angelova and S. Zhu, "Efficient object detection and segmentation for fine-grained recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 811–818.
- [57] C. Huang, H. Li, Y. Xie, Q. Wu, and B. Luo, "PBC: Polygon-based classifier for fine-grained categorization," *IEEE Trans. Multimedia*, vol. 19, no. 4, pp. 673–684, Apr. 2017.
- [58] L. Xie, J. Wang, W. Lin, B. Zhang, and Q. Tian, "Towards reversal-invariant image representation," *Int. J. Comput. Vis.*, vol. 123, no. 2, pp. 226–250, 2017.
- [59] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPR)*, Jun. 2014, pp. 806–813.
- [60] H. Azizpour and I. Laptev, "Object detection using strongly-supervised deformable part models," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 836–849.
- [61] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
- [62] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. (May 2014). "Return of the devil in the details: Delving deep into convolutional nets." [Online]. Available: <https://arxiv.org/abs/1405.3531>



Yuxin Peng is the professor of Institute of Computer Science and Technology (ICST), Peking University, Beijing, China, and the chief scientist of 863 Program (National Hi-Tech Research and Development Program of China). He received the Ph.D. degree in computer application from Peking University, in July 2003. After that, he worked as an assistant professor in ICST, Peking University. He was promoted to an associate professor and professor in Peking University in August 2005 and August 2010, respectively.

In 2006, he was authorized by the Program for New Star in Science and Technology of Beijing and the Program for New Century Excellent Talents in University (NCET). He has published more than 100 papers in refereed international journals and conference proceedings, including IJCV, TIP, TMM, TCSVT, PR, ACM MM, ICCV, CVPR, IJCAI, AAAI, etc. He led his team to participate in TRECVID (TREC Video Retrieval Evaluation) many times. In TRECVID 2009, his team won four first places on 4 sub-tasks of the High-Level Feature Extraction task and Search task. In TRECVID 2012, his team gained four first places on all 4 sub-tasks of the Instance Search (INS) task and Known-Item Search task. In TRECVID 2014, his team gained the first place in the Interactive Instance Search task. His team also gained both two first places in the INS task of TRECVID 2015, 2016, and 2017. Besides, he won the first prize of Beijing Science and Technology Award in 2016 (ranking first). He has applied 35 patents, and obtained 16 of them. His current research interests mainly include cross-media analysis and reasoning, image and video analysis and retrieval, and computer vision.



Xiangteng He received the B.S. degree in computer science and technology from Nankai University in 2014. He is currently pursuing the Ph.D. degree with the Institute of Computer Science and Technology, Peking University. His current research interests include image analysis and deep learning.



Junjie Zhao received the B.S. degree in computer science and technology from Peking University in 2015, where he is currently pursuing the M.S. degree with the Institute of Computer Science and Technology. His current research interests include image analysis and deep learning.