# Fast Fine-Grained Image Classification via Weakly Supervised Discriminative Localization

Xiangteng He, Yuxin Peng, and Junjie Zhao

***Abstract*— Fine-grained image classification is to recognize hundreds of subcategories in each basic-level category. Existing methods employ discriminative localization to find the key distinctions between similar subcategories. However, they generally have two limitations: 1) discriminative localization relies on region proposal methods to hypothesize the locations of discriminative regions, which are *time-consuming* and the *bottleneck* of improving classification speed and 2) the training of discriminative localization depends on object or part annotations which are heavily *labor-consuming* and the *obstacle* of marching toward practical application. It is highly challenging to address the two limitations *simultaneously*, while existing methods only focus on one of them. Therefore, we propose a weakly supervised discriminative localization approach (WSDL) for fast fine-grained image classification to address the two limitations at the same time, and its main advantages are: 1) multi-level attention guided localization learning is proposed to localize discriminative regions with different focuses automatically, without using object and part annotations, avoiding the labor consumption. Different level attentions focus on different characteristics of the image, which are complementary and boost classification accuracy and 2) *n*-pathway end-to-end discriminative localization network is proposed to improve classification speed, which simultaneously localizes multiple different discriminative regions for one image to boost classification accuracy, and shares full-image convolutional features generated by a region proposal network to accelerate the process of generating region proposals as well as reduce the computation of convolutional operation. Both are jointly employed to simultaneously improve classification speed and eliminate dependence on object and part annotations. Comparing with state-of-the-art methods on two widely used fine-grained image classification data sets, our WSDL approach achieves the best accuracy and the efficiency of classification.**

***Index Terms*— Fast fine-grained image classification, weakly supervised discriminative localization, multi-level attention.**

## I. Introduction

FINE-GRAINED image classification aims to recognize hundreds of subcategories in the same basic-level category, which lies in the continuum between basic-level image classification (e.g. object recognition [2], [3]) and identification of individuals (e.g. face recognition [4], [5]). It is one of the most significant and highly challenging open problems in computer vision area due to the following two aspects: (1) *Large variance in the same subcategory.* As shown in the first row of Fig. 1, the four images belong to the same subcategory of "Laysan Albatross", but they are different in poses, views, feathers and so on. It is easy for human beings to misclassify them into different subcategories. (2) *Small variance among different subcategories.* As shown in the second row of Fig. 1, the four images belong to different subcategories, but they are all black and look similar. It is hard for human beings to distinguish "Fish Crow" from the other three subcategories. These subcategories in the same basic-level category look similar in global appearance, but distinct in some discriminative regions of the objects, such as the head. So the localization of the key discriminative regions becomes crucial for fine-grained image classification. Recently, methods based on discriminative localization have achieved great progress [6]–[12].

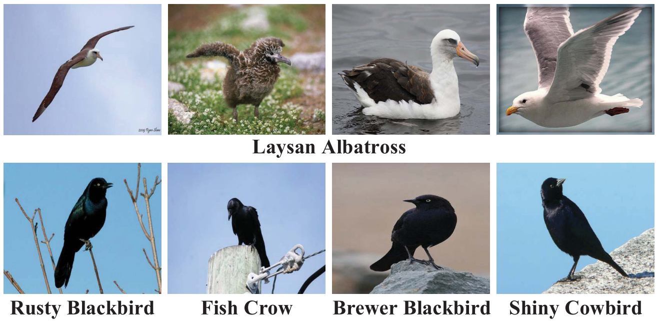

Fig. 1. Examples of CUB-200-2011 dataset [1]. Large variance in the same subcategory is shown in the first row, and small variance among different subcategories is shown in the second row.

Fine-grained image classification has wide applications in automatic driving, biological conservation, cancer detection, and so on. In the process of converting technology into application, there are two important problems that need to be solved urgently: (1) *Time consumption.* Some existing methods mainly focus on achieving better classification accuracy, but ignore the problem of time consumption. However, real-time performance is one of the most important criteria in the application of fine-grained image classification, which satisfies the response speed requirements of users. (2) *Labor consumption.* The annotations of image (e.g. the image-level subcategory label, the bounding box of the object and part locations) are required in the training phase of many existing methods, and even in the testing phase. While the annotations are labor-consuming and unrealistic in the applications of fine-grained

Manuscript received September 30, 2017; revised February 15, 2018; accepted May 4, 2018. Date of publication May 8, 2018; date of current version May 3, 2019. This work was supported by the National Natural Science Foundation of China under Grant 61771025 and Grant 61532005. This paper was recommended by Associate Editor A. Loui. *(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/TCSVT.2018.2834480

image classification. So utilizing as few annotations as possible is the key point to convert fine-grained image classification into application.

Early works only focus on achieving better classification accuracy, but ignore aforementioned two problems. Zhang *et al.* [6] propose the Part-based R-CNN method, which learns whole-object and part detectors with geometric constraints between them. The learning phase of detectors depends on the annotations of image-level subcategory label, object and part. They first generate thousands of region proposals for each image via Selective Search method [13], which is one of the most popular region proposal methods, and greedily merges pixels based on engineered low-level features. Then they utilize the learned whole-object and part detectors to detect object and parts from the generated region proposals, and finally predict a fine-grained subcategory based on a pose-normalized representation. This framework is widely used in fine-grained image classification. Krause *et al.* [8] adopt the box constraint of Part-based R-CNN [6] to train part detectors with only object annotations. These methods rely on region proposal methods implemented with CPU to hypothesize the locations of discriminative regions, which are *time-consuming* and the *bottleneck* of improving classification speed. Discriminative localization learning depends on object or part annotations, which are heavily *labor-consuming* and the *obstacle* of marching towards practical application. It is highly challenging and significant to address these two problems *simultaneously*, while existing methods only focus on one of them.

*For addressing the problem of time consumption*, researchers focus on designing end-to-end network and avoiding the application of the time-consuming region proposal methods implemented with CPU. Zhang *et al.* [14] propose the Part-stacked CNN architecture, which consists of a fully convolutional network and a two-stream classification network. They first utilize fully convolutional network to localize discriminative regions, and then adopt the two-stream classification network to encode object-level and part-level features simultaneously. Part-stacked CNN is over two orders of magnitude faster than Part-based R-CNN [6], but requires the annotations of image-level subcategory label, object and part in the training phase, which is *labor-consuming*.

*For addressing the problem of labor consumption*, researchers focus on the localization of the discriminative regions under weakly supervised setting, which denotes that neither object nor part annotations are used in both training and testing phases. Xiao *et al.* [9] propose the two-level attention model: object-level attention is to select region proposals relevant to a certain object, and part-level attention is to localize discriminative parts. It is the first work to classify fine-grained images without using object or part annotations in both training and testing phases, but still achieve promising results [15]. Simon and Rodner [11] propose a constellation model to localize discriminative regions of object, leveraging CNN to find the constellations of neural activation patterns. A part model is estimated by selecting part detectors via constellation model. And then the part model is used to extract features for classification. These methods avoid depending on object or part annotations, but they generally utilize Selective Search [13] method to generate region proposals, which is *time-consuming*.

Existing methods only focus on solving one of *labor consumption* and *time consumption* problems, achieving improvement at the sacrifice of the other one, which may cause that the other problem becomes worse. Therefore, this paper proposes a *weakly supervised discriminative localization approach (WSDL) for fast fine-grained image classification*, which aims to *simultaneously* solve the above two problems, improving classification speed and eliminating dependence on object and parts annotations at the same time. Its main contributions can be summarized as follows:

- ***Multi-level attention guided localization learning.*** Existing weakly supervised localization methods directly utilize attention maps to generate discriminative regions, which have two limitations: slow localization speed and low classification accuracy. Therefore, we propose a new multi-level attention guided localization learning approach to implement localization and classification simultaneously. Attention maps are applied to guide the secondary localization learning for more accurate localization and faster localization speed, avoiding the cost of processing of attention maps in existing weakly supervised localization methods. Different level attentions describe the visual content at different characteristics, carrying multi-grained and multi-scale information. They are complementary to each other for boosting classification accuracy. The learning process is guided by attention maps, without using object and part annotations, avoiding the labor consumption.
- ***n-pathway end-to-end discriminative localization network.*** Existing localization methods localize only one discriminative region at one time, ignoring other discriminative regions, which should be considered to boost the classification accuracy. Therefore, we propose a new n-pathway end-to-end discriminative localization network to localize different discriminative regions for an image at the same time. It consists of multiple localization networks and one region proposal network. Multiple localization networks share full-image convolutional features generated by region proposal network, to reduce the computation of convolutional operation, and avoid the nearly linear growth of time consumption caused by the localization of multiple discriminative regions.

Our previous conference paper [16] proposes a discriminative localization approach via saliency-guided Faster R-CNN, which localizes the discriminative region in the image to boost the classification accuracy. The main differences between the proposed WSDL approach and our previous conference paper [16] can be summarized as the following two aspects: (1) Our previous conference paper [16] applies one level attention, while our WSDL approach further employs multi-level attention to guide the discriminative localization learning, which localizes multi-grained and multi-scaled discriminative regions to boost fine-grained classification accuracy. (2) Our WSDL approach designs $n$-pathway network structure to reduce the growth of time consumption in classification.

Time consumption is reduced by sharing full-image convolutional features among different localization networks with different level attentions. The architecture in [16] can only deal with one level attention, and the application of multi-level attention will cause the nearly linear growth of time consumption in classification. Comparing with state-of-the-art methods on two widely-used fine-grained image classification datasets, the effectiveness of our WSDL approach is verified, achieving both the best accuracy and efficiency of classification.

The rest of this paper is organized as follows: Section II briefly reviews related works on fine-grained image classification and object detection, Section III presents our WSDL approach in detail, and Section IV introduces the experimental results as well as the experimental analyses. Finally Section V concludes this paper and presents the future works.

## II. Related Work

In this section, we review related works on fine-grained image classification and object detection.

### *A. Fine-Grained Image Classification*

Fine-grained image classification is one of the most fundamental and challenging open problems in computer vision, and has drawn extensive attention in both academia and industry. Early works [17], [18] focus on the design of feature representation and classifier based on the basic low-level descriptors, such as SIFT [19]. The performance of these methods is limited due to the handcrafted features. Recently, deep learning has achieved great success in computer vision, speech recognition, natural language processing, and so on. Inspired by this, many researchers begin to study on the problem of fine-grained image classification by deep learning [6], [9], [10], [12], [15], and have achieved great progress.

Since discriminative characteristics generally localize in the regions of object and its parts, most existing works generally follow the two-stage pipeline: First localize the object and parts, and then extract their features to train classifiers. For the first stage, some works [20], [21] directly utilize the human annotations (i.e. the bounding box of the object and part locations) to localize the object and parts. Since the human annotations are labor-consuming, some researchers begin to only utilize them in the training phase. Zhang *et al.* [6] propose the Part-based R-CNN to directly utilize the object and part annotations to learn the whole-object and part detectors with geometric constraints between them. This framework is widely used in fine-grained image classification.

Recently, fine-grained image classification methods begin to focus on how to achieve promising performance without using any object or part annotations. The first work under such weakly supervised setting is the two-level attention model [9], which utilizes the attention mechanism of the CNNs to select region proposals corresponding to the object and parts, and achieves promising results even compared with those methods relying on the object and part annotations. Inspired by this work, Zhang *et al.* [10] incorporate deep convolutional filters for both parts selection and description. He and Peng [12] integrate two spatial constraints for improving the performance of parts selection.

### *B. Object Detection*

Object detection is one of the most fundamental and challenging open problems in computer vision, which not only recognizes the objects but also localizes them in the images. Like fine-grained image classification, early works are mainly based on basic low-level features, such as SIFT [19] and HOG [22]. However, since 2010, the progress of handcrafted features based object detection slows down. Due to the great success of deep learning in the competition of ImageNet LSVRC-2012, deep learning has been widely employed in computer vision, including object detection. We divide the CNN-based object detection methods into two groups by the annotations used: (1) Supervised object detection, which needs ground-truth bounding box of the object. (2) Weakly supervised object detection, which does not need ground-truth bounding box of the object, and only needs image-level labels.

*1) Supervised Object Detection:* Girshick *et al.* [23] propose a simple and scalable detection framework, regions with CNN features, called R-CNN. First, it utilizes the region proposal method (i.e. Selective Search [13]) to generate thousands of region proposals for each image. Then it trains CNNs end-to-end to extract highly discriminative features of these region proposals. Finally, it classifies these region proposals based on their discriminative features to determine whether the region proposal can be output as a bounding box of the object in the image. Inspired by R-CNN, many works follow the pipeline: First utilize the region proposal methods to generate region proposals for each image, and then employ CNNs to extract their features and classify their category.

However, these methods have a limitation: Time consumption is high because each region proposal needs to pass forward CNNs respectively, while each image generally generates thousands of region proposals. This limitation causes that object detection cannot satisfy the requirement of real-time performance. For addressing this limitation, SPP-net [24] and Fast R-CNN [3] are proposed. SPP-net applies a spatial pyramid pooling (SPP) layer to pool a fixed-length feature representation of each region proposal, which extracts the feature maps from the entire image only once, and avoids the time-consuming convolutional operation of each region proposal. Comparing with R-CNN, SPP-net is $24 \sim 102 \times$ faster in object detection. However, SPP-net cannot update the convolutional layers before the spatial pyramid pooling layer, and its extracted features need to be stored to disk, which limits both the accuracy and efficiency. Therefore, Fast R-CNN [3] is proposed to fix the disadvantages of R-CNN and SPP-net. Fast R-CNN utilizes a region of interest (RoI) pooling layer to extract a fixed-length feature vector for each region proposal based on the feature map, employs multi-task loss to train the network in a single-stage, and updates all network layers in the training phase. Comparing with SPP-net, Fast R-CNN is 10 × faster and more accurate in object detection.

However, all the above methods are based on the region proposal methods, such as Selective Search [13], EdgeBoxes [25], which become the computational bottleneck. These region proposal methods are implemented with CPU, which causes that the time consumption of generating region proposals

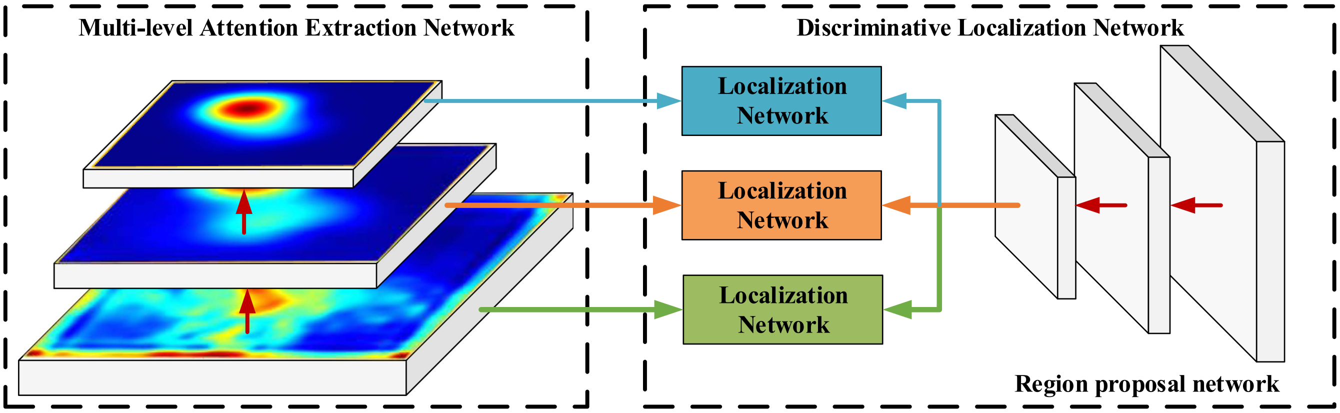

Fig. 2. An overview of our WSDL approach. Multi-level attention extraction network (MAEN) extracts the attention information from multiple convolutional layers to provide the bounding boxes of discriminative regions for training discriminative localization network (DLN). The DLN consists of one region proposal network and multiple localization networks. We utilize 3-level attention and 3 localization networks in the figure to clearly demonstrate our WSDL approach.

is high. Therefore, Faster R-CNN [26] proposes a region proposal network (RPN) to generate region proposals and implements it with GPU, which makes the computation of generating region proposals nearly cost-free.

*2) Weakly Supervised Object Detection:* All the above object detection methods need ground-truth bounding box of the object in the training phase, which is labor-consuming. Recently, many methods [27]–[29] have begun to exploit weakly supervised object detection based on CNNs. These methods mainly base on activation maps of convolutional layer, which is widely named as attention maps, such as in [29]–[32]. As described in [30], attention maps are considered as a set of spatial maps that essentially try to encode on which spatial areas of the input the network focuses most for taking its output decision (e.g., for classifying an image), where, furthermore, these maps can be defined w.r.t. various layers of the network so that they are able to capture both low-, mid-, and high-level representation information. Oquab *et al.* [27] propose a weakly supervised learning method based on an end-to-end CNN only with image-level labels, and utilize max pooling operation to generate the attention map to localize the object. Zhou *et al.* [29] propose CAM, which uses global average pooling (GAP) in CNN to generate the attention map for each image. Based on the attention map, the region of the object can be localized. Inspired by CAM [29], we remove the fully-connected layers before final output in CNN and replace them with global average pooling followed by a fully-connected softmax layer, producing attention maps of different layers for each subcategory. From the attention maps, we can obtain the discriminative regions that convolutional layers attend to, and they are significant for the classification.

## III. Weakly Supervised Discriminative Localization

We propose *a weakly supervised discriminative localization approach (WSDL) for fast fine-grained image classification*, where an $n$-pathway end-to-end network is designed to localize discriminative regions and encode discriminative features simultaneously. Despite achieving a notable classification accuracy, our WSDL approach improves classification speed as well as eliminates dependence on object and part annotations simultaneously. An overview of our approach is shown as Fig. 2. It consists of two subnetworks: multi-level attention extraction network (MAEN) and discriminative localization network (DLN). Multi-level attention extraction network extracts multi-level attention information from different convolutional layers for each image, and generates multiple initialized discriminative regions based on the attention information. Then the bounding boxes of these discriminative regions are adopted as the annotations to guide the training of discriminative localization network, which localizes multiple discriminative regions, avoiding the dependence on object and part annotations. Both MAEN and DLN can generate the discriminative regions, but with different advantages: (1) Instead of using the labor-consuming human annotations, MAEN provides the bounding box information of discriminative regions for the training of DLN automatically, even though the discriminative region is not very accurate. It is noted that MAEN is only employed in the training phase. (2) Based on the initialized discriminative regions generated by MAEN, DLN further optimizes the learned discriminative regions to find where are more discriminative for distinguishing this subcategory from others. Their combination makes the best of their advantages and fixes their disadvantages to further achieve better classification performance.

### *A. Multi-Level Attention Extraction Network*

Attention is a behavioral and cognitive process of selectively concentrating on a discrete aspect of information [33]. Tsotsos *et al.* [34] state that visual attention mechanism seems to involve the selection of regions of interest in the visual field. And Karklin and Lewicki [35] indicate that neurons in primary visual cortex (e.g. V1) respond to the edge over a range of positions, and neurons in higher visual areas (e.g. V2 and V4) are more invariant to image properties and might encode shape. The discovery is also shown in convolutional neural networks (CNNs), different feature maps (attention) reflect different characteristics of the image. The images in different rows of Fig. 3 are the attention maps extracted from the

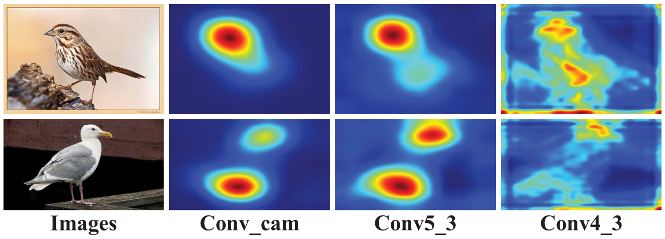

Fig. 3. Examples of attention maps extracted by MAEN in our WSDL approach.

convolutional layers of "Conv4_3", "Conv5_3" and "Conv_cam" in our multi-level attention extraction network respectively. We can observe that different convolutional layers have different focuses, and provide complementary information to boost the classification accuracy.

According to the studies on visual attention mechanism, we design the multi-level attention extraction network (MAEN) to generate bounding box information for discriminative localization network. We take resized images as inputs and output $n$ feature maps from $n$ convolutional layers as the multi-level attention maps to indicate the importance of each pixel in the image for classification. Then we generate the bounding boxes of discriminative regions based on these attention maps. Inspired by CAM [29], we remove the fully-connected layers before final output in CNN and replace them with global average pooling followed by a fully-connected softmax layer. Then we sum the feature maps of the certain convolutional layer with weights to generate attention map for each image. In this stage, we will generate $n$ attention maps based on $n$ different convolutional layers. Finally, we perform binarization operation on each attention map with an adaptive threshold, which is obtained by OTSU algorithm [36], and take the bounding box that covers the largest connected area as the discriminative region. $n$ bounding boxes of discriminative regions are adopted as bounding box information of discriminative localization network.

For a given image $I$, we generate $n$ attention maps, the value of spatial location $(x, y)$ in $i$-th attention map is defined as follow:

$$M_i(x, y) = \sum_{u_i} w_{u_i} f_{u_i}(x, y) \tag{1}$$

where $M_i(x, y)$ indicates the importance of activation at spatial location $(x, y)$ for classification, $f_{u_i}(x, y)$ denotes the activation of neuron $u_i$ in the $i$-th convolutional layer at spatial location $(x, y)$, and $w_{u_i}$ denotes the weight used to sum the activation $f_{u_i}(x, y)$ to generate the attention map. For different convolutional layers, $w_{u_i}$ has different definitions as follow:

$$w_{u_i} = \begin{cases} w_{u_i}^c, & last\ convolutional\ layer \\ \dfrac{1}{|u_i|}, & otherwise \end{cases} \tag{2}$$

where $w_{u_i}^c$ denotes the weight corresponding to subcategory $c$ for neuron $u_i$ in the last convolutional layer, denoted as "Conv_cam" in our MAEN. We use the predicted result as the subcategory $c$ instead of using the image-level subcategory label. $|u_i|$ denotes the total number of neurons in $i$-th convolutional layer.

### B. Discriminative Localization Network

From multi-level attention extraction network, we obtain $n$ attention maps to guide the training of discriminative localization network. To make the best of the complementarity of multi-level attention information, we design an $n$-pathway end-to-end network based on Faster R-CNN [26], which consists of multiple localization networks and one region proposal network. Faster R-CNN is proposed to accelerate the process of detection as well as achieve promising detection performance. We modify the original Faster R-CNN in two aspects: (1) The training phase of Faster R-CNN needs ground-truth bounding box of discriminative region in the image, which is heavily labor-consuming. In this paper, we use the bounding box information provided by multi-level attention extraction network as the ground-truth bounding box information, which eliminates the dependence on object and part annotations. (2) Inspired by the discoveries on visual attention mechanism, we apply multi-level attention into our WSDL approach. However, the application of multi-level attention is restricted by the architecture of the original Faster R-CNN. Original Faster R-CNN consists of one region proposal network and one localization network, which restricts it to only localize one discriminative region at one time. We need to train $n$ Faster R-CNN models to apply the multi-level attention, which causes the nearly linear growth of time consumption in classification. Therefore, we design an $n$-pathway end-to-end network with multiple localization networks and one region proposal network, where all the localization networks share the same full-image convolutional features generated by region proposal network.

Instead of using time-consuming region proposal methods such as Selective Search method [13], region proposal network (RPN) is designed to quickly generate region proposals for each image by sliding a small network over the feature map of last shared convolutional layer. At each sliding-window location, $k$ region proposals are simultaneously predicted, and they are parameterized relative to $k$ anchors. We apply 9 anchors with 3 scales and 3 aspect ratios. For training RPN, a binary class label of being an object or not is assigned to each anchor, which depends on the Intersection-over-Union (IoU) [37] overlap with a ground-truth bounding box of the object. But in our WSDL approach, we compute the IoU overlap with the bounding boxes of discriminative regions generated by MAEN rather than the ground-truth bounding box of the object. And the loss function for an image is defined as:

$$L(\{p_i\}, \{t_i\}) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*) \tag{3}$$

where $i$ denotes the index of an anchor in a mini-batch, $p_i$ denotes the predicted probability of anchor $i$ being a discriminative region, $p_i^*$ denotes the label of being a discriminative region of object or not depending on the bounding box $t_i^*$ generated by MAEN, $t_i$ is the predicted bounding box of discriminative region, $L_{cls}$ is the classification loss defined by log loss, and $L_{reg}$ is the regression loss defined by the robust loss function (smooth $L_1$).

Since we apply multi-level attention into our WSDL approach, we employ $n$ localization networks, and each of them is the same with Fast R-CNN [3]. All the localization networks are connected to RPN by a region of interest (RoI) pooling layer, which is employed to extract a fixed-length feature vector from feature map for each region proposal generated by RPN. Each feature vector is taken as the input of each localization network, and passes forward to generate two outputs: one is predicted subcategory and the other is predicted bounding box of discriminative region. Through discriminative localization network, we obtain the discriminative regions with multi-level attention. Then we average the predicted scores of the discriminative regions with multi-level attention and the original image to obtain the subcategory of each image.

### *C. Training of MAEN and DLN*

MAEN learns the multi-level attention information of image to tell which regions are important and discriminative for classification, and then guides the training of DLN. RPN in DLN generates region proposals relevant to the discriminative regions of images. Considering that training RPN needs bounding boxes of discriminative regions provided by MAEN, we adopt the strategy of sharing convolutional weights between MAEN and RPN to promote the localization learning.

First, we train the MAEN. This network is first pre-trained on the ImageNet 1K dataset [38], and then fine-tuned on the fine-grained image classification dataset, such as CUB-200-2011 dataset [1]. Second, we train the RPN. Its initial weights of convolutional layers are cloned from MAEN. Instead of fixing the shared convolutional layers, all layers are fine-tuned in the training phase. Third, we train the localization networks. Since all the localization networks share full-image convolutional features generated by RPN, we fix the parameters of RPN when training the localization networks with multi-level attention.

### *D. Implementation Details*

Our WSDL approach consists of two subnetworks: multi-level attention extraction network (MAEN) and discriminative localization network (DLN). Both of them are all based on 16-layer VGGNet [39], which is widely used in image classification task. It can be replaced with the other CNNs. MAEN extracts the attention information of images to provide bounding boxes needed by DLN. For VGGNet in MAEN, we remove the layers after “Conv5_3” and add a convolutional layer of size 3 × 3, stride 1, pad 1 with 1024 neurons, which is followed by a GAP layer and a softmax layer [29]. We adopt the object-level attention of Xiao *et al.* [9] to select relevant image patches for data extension. And then we utilize the extended data to fine-tune MAEN for learning discriminative features. The number of neurons in softmax layer is set as the number of subcategories in the dataset. DLN shares the weights of layers before “Conv5_3” with MAEN for better discriminative localization as well as classification performance.

At training phase, for MAEN, we initialize the weights with the network pre-trained on the ImageNet 1K dataset [38], and then use SGD with a minibatch size of 20. We use a weight decay of 0.0005 with a momentum of 0.9 and set the initial learning rate as 0.001. The learning rate is divided by 10 every 5K iterations. We terminate the training phase at 35K iterations on CUB-200-2011 dataset [1] and 55K iterations on Cars-196 dataset [40] because of different convergence rate. The discriminative localization network designed in our WSDL approach consists of one RPN and $n$ localization networks. In the training phase, each localization network is trained one by one with RPN. We first initialize the weights of the convolutional layers in RPN with the MAEN, and then train the RPN and 3 localization networks. When training the localization networks, the weights of RPN are fixed, and only the weights of the localization network are fine-tuned. For the training of RPN and localization network, we start SGD with a minibatch size of 128, use a weight decay of 0.0005 with a momentum of 0.9 and set the initial learning rate to 0.001. We divide the learning rate by 10 at 40K iterations on CUB-200-2011 dataset and 50K iterations on Cars-196 dataset, and terminate training at 90K iterations on CUB-200-2011 dataset and 120K iterations on Cars-196 dataset.

## IV. Experiments

We conduct experiments on 2 widely-used datasets in the fine-grained image classification task: CUB-200-2011 [1] and Cars-196 [40] datasets. Our WSDL approach is compared with state-of-the-art methods to verify its effectiveness, where our WSDL approach achieves both the best accuracy and efficiency of fine-grained image classification.

### *A. Datasets*

Two datasets are adopted in the experiments:

- **CUB-200-2011** [1] is the most widely-used dataset in fine-grained image classification task, which contains 11788 images of 200 subcategories belonging to the category of bird, 5994 images in the training set and 5794 images in the testing set. Each image is labeled with detailed annotations including an image-level subcategory label, a bounding box of the object and 15 part locations. In our experiments, only image-level subcategory label is used in the training phase.
- **Cars-196** [40] contains 16185 images of 196 car subcategories, which is divided as follows: the training set contains 8144 images, and the testing set contains 8041 images. For each subcategory, 24∼84 images are selected for training and 24∼83 images for testing. Every image is annotated with an image-level subcategory label and a bounding box of the object. The same with CUB-200-2011 dataset, only image-level subcategory label is used in the training phase.

### *B. Evaluation Metrics*

**Accuracy** is adopted to comprehensively evaluate the classification performances of our WSDL approach as well as the

TABLE I

COMPARISONS WITH STATE-OF-THE-ART METHODS ON CUB-200-2011 DATASET IN THE ASPECT OF CLASSIFICATION ACCURACY TO SHOW THE EFFECTIVENESS OF OUR WSDL APPROACH

| Methods | Train Annotation | | Test Annotation | | Accuracy (%) | CNN Features |
|---|---|---|---|---|---|---|
| | Object | Parts | Object | Parts | | |
| **Our WSDL Approach** | | | | | **85.71** | VGGNet |
| Saliency-guided Faster R-CNN [16] | | | | | 85.14 | VGGNet |
| TSC [12] | | | | | 84.69 | VGGNet |
| FOAF [41] | | | | | 84.63 | VGGNet |
| PD [10] | | | | | 84.54 | VGGNet |
| STN [42] | | | | | 84.10 | GoogleNet |
| Bilinear-CNN [43] | | | | | 84.10 | VGGNet&VGG-M |
| PD (FC-CNN) [10] | | | | | 82.60 | VGGNet |
| Multi-grained [44] | | | | | 81.70 | VGGNet |
| NAC [11] | | | | | 81.01 | VGGNet |
| PIR [15] | | | | | 79.34 | VGGNet |
| RBF [45] | | | | | 78.98 | ResNet-50 |
| TL Atten [9] | | | | | 77.90 | VGGNet |
| MIL [46] | | | | | 77.40 | VGGNet |
| VGG-BGLm [47] | | | | | 75.90 | VGGNet |
| InterActive [48] | | | | | 75.62 | VGGNet |
| Coarse-to-Fine [49] | √ | | √ | | 82.90 | VGGNet |
| PG Alignment [8] | √ | | √ | | 82.80 | VGGNet |
| Coarse-to-Fine [49] | √ | | | | 82.50 | VGGNet |
| VGG-BGLm [47] | √ | | √ | | 80.40 | VGGNet |
| Triplet-A (64) [50] | √ | | √ | | 80.70 | GoogleNet |
| Triplet-M (64) [50] | √ | | √ | | 79.30 | GoogleNet |
| AGAL [51] | | √ & attribute | | | 85.40 | ResNet-50 |
| Webly-supervised [52] | √ | √ | | | 78.60 | AlexNet |
| PN-CNN [53] | √ | √ | | | 75.70 | AlexNet |
| Part-based R-CNN [6] | √ | √ | | | 73.50 | AlexNet |
| AGAL [51] | √ | √ & attribute | | | 85.50 | ResNet-50 |
| SPDA-CNN [14] | √ | √ | √ | | 85.14 | VGGNet |
| Deep LAC [54] | √ | √ | √ | | 84.10 | AlexNet |
| SPDA-CNN [14] | √ | √ | √ | | 81.01 | AlexNet |
| Part-stacked CNN [55] | √ | √ | √ | | 76.20 | AlexNet |
| PN-CNN [53] | √ | √ | √ | √ | 85.40 | AlexNet |
| Part-based R-CNN [6] | √ | √ | √ | √ | 76.37 | AlexNet |
| POOF [21] | √ | √ | √ | √ | 73.30 | |
| HPM [20] | √ | √ | √ | √ | 66.35 | |

compared state-of-the-art methods, which is widely used in fine-grained image classification [6], [10], [15], and defined as follow:

$$Accuracy = \frac{R_a}{R} \tag{4}$$

where $R$ denotes the number of images in the testing set (e.g. $R$ equals to 5794 in the CUB-200-2011 dataset), and $R_a$ denotes the number of images that are correctly classified.

**Intersection-over-Union (IoU)** [37] is adopted to evaluate whether the predicted bounding box of discriminative region is a correct localization, and its definition is as follow:

$$IoU = \frac{area(B_p \cap B_{gt})}{area(B_p \cup B_{gt})} \tag{5}$$

where $B_p$ denotes the predicted bounding box of discriminative region, $B_{gt}$ denotes the ground-truth bounding box of the object, $B_P \cap B_{gt}$ denotes the intersection of the predicted and ground-truth bounding boxes, and $B_p \cup B_{gt}$ denotes their union. The predicted bounding box of discriminative region is correctly localized, if the $IoU$ exceeds 0.5.

### *C. Comparisons With State-of-the-Art Methods*

This subsection presents the experimental results and analyses of our WSDL approach as well as the compared state-of-the-art methods on the widely-used CUB-200-2011 [1] and Cars-196 [40] datasets. We verify the effectiveness of our WSDL approach in the following aspects: (1) Accuracy of classification. (2) Efficiency of classification. The experimental results show that our WSDL achieves better performance than state-of-the-art methods in both accuracy and efficiency of classification.

*1) Accuracy of Classification:* Tables I and II show the comparisons with state-of-the-art methods on CUB-200-2011 and Cars-196 datasets in the aspect of classification accuracy. Object, part annotations and CNN features used in these methods are listed for fair comparison. As object annotation is easily confused with image-level subcategory label, we explain as follows: Image-level subcategory label is different from object annotation. Image-level label denotes which subcategory the image belongs to, such as “Laysan Albatross” and “Rusty Blackbird”. While object level label denotes bounding box of the object in an image, which is an axis-aligned rectangle specifying the extent of the object. The bounding box is represented as $(x, y, width, height)$ , where $(x, y)$ denotes the coordinate information of the top left corner of the object, and $(width, height)$ denotes the width and height of the object. In Tables I and II, “Object” means ground-truth bounding box, not image-level label. Our WSDL approach

TABLE II

COMPARISONS WITH STATE-OF-THE-ART METHODS ON CARS-196 DATASET IN THE ASPECT OF CLASSIFICATION ACCURACY TO SHOW THE EFFECTIVENESS OF OUR WSDL APPROACH

| Methods | Train Annotation | | Test Annotation | | Accuracy (%) | CNN Features |
|---|---|---|---|---|---|---|
| | Object | Parts | Object | Parts | | |
| **Our WSDL Approach** | | | | | **92.30** | VGGNet |
| Saliency-guided Faster R-CNN [16] | | | | | 91.87 | VGGNet |
| Bilinear-CNN [43] | | | | | 91.30 | VGGNet&VGG-M |
| TL Atten [9] | | | | | 88.63 | VGGNet |
| DVAN [56] | | | | | 87.10 | VGGNet |
| FT-HAR-CNN [57] | | | | | 86.30 | AlexNet |
| HAR-CNN [57] | | | | | 80.80 | AlexNet |
| PG Alignment [8] | √ | | | | 92.60 | VGGNet |
| SWP [58] | √ | | | | 92.30 | ResNet-50 |
| ELLF [59] | √ | | | | 73.90 | CNN |
| R-CNN [23] | √ | | | | 57.40 | AlexNet |
| PG Alignment [8] | √ | | √ | | 92.80 | VGGNet |
| BoT(CNN With Geo) [60] | √ | | √ | | 92.50 | VGGNet |
| DPL-CNN [61] | √ | | √ | | 92.30 | VGGNet |
| VGG-BGLm [47] | √ | | √ | | 90.50 | VGGNet |
| BoT(HOG With Geo) [60] | √ | | √ | | 85.70 | VGGNet |
| LLC [62] | √ | | √ | | 69.50 | |
| BB-3D-G [40] | √ | | √ | | 67.60 | |

TABLE III

COMPARISONS WITH STATE-OF-THE-ART METHODS IN THE ASPECT OF CLASSIFICATION EFFICIENCY. CUB-200-2011 DATASET IS ADOPTED AS THE EVALUATION DATASET, AND AVERAGE CLASSIFICATION SPEED IS EVALUATED BY THE FRAMES RECOGNIZED PER SECOND, DENOTED AS FPS. THE RESULTS ARE ALL OBTAINED ON THE COMPUTER WITH ONE GPU OF NVIDIA TITAN X @1417MHZ AND ONE CPU OF INTEL CORE I7-6900K @3.2GHZ

| Methods | Average Classification Speed (fps) | CNN Models |
|---|---|---|
| **Our WSDL Approach** | **9.09** | VGGNet |
| Saliency-guided Faster R-CNN [16] | 10.07 | VGGNet |
| Bilinear-CNN [43] | 4.52 | VGGNet&VGG-M |
| TSC [12] | 0.34 | VGGNet |
| TL Atten [9] | 0.25 | VGGNet |
| NAC [11] | 0.10 | VGGNet |
| **Our WSDL Approach** | **16.13** | AlexNet |
| Saliency-guided Faster R-CNN [16] | 17.09 | AlexNet |
| Part-stacked CNN [55] | 14.30 | AlexNet |

only uses image level labels, which avoids the heavy labor consumption of the labeling of bounding box, but still achieves the best classification accuracy. CNN model shown in the column of “CNN Features” indicates which CNN model is adopted to extract features. If the method adopt handcrafted feature like SIFT, the column of “CNN Features” is none. We present detailed analyses of our WSDL approach as well as compared methods on CUB-200-2011.

Early methods choose SIFT [19] as basic feature and even use both object and part annotations, such as POOF [21] and HPM [20], but their classification results are limited and much lower than our WSDL approach. Our WSDL approach achieves the highest classification accuracy among all the state-of-the-art methods under the same weakly supervised setting, which indicates that neither object nor part annotations are used both in training and testing phases. Our WSDL achieves the improvement by 1.02% than the best state-of-the-art result of TSC [12] (85.71% vs. 84.69%), which jointly considers two spatial constraints in part selection. Despite achieving better classification accuracy, our WSDL approach is over two order of magnitude faster (i.e. 27 × faster) than TSC, as shown in Table III. The efficiency analyses will be described latter. And our WSDL approach achieves better classification accuracy than the method of Bilinear-CNN [43], which combines two different CNNs: VGGNet [39] and VGG-M [63]. Its classification accuracy is 84.10%, lower than our approach by 1.61%.

Even compared with state-of-the-art methods using object annotations in both training and testing phases, such as Coarse-to-Fine [49], PG Alignment [8] and VGG-BGLm [47], our WSDL approach achieves improvement by at least 2.81%. Moreover, our WSDL approach outperforms state-of-the-art methods using both object and part annotations, such as SPDA-CNN [14]. Neither object nor part annotations are used in our WSDL approach, which marches toward practical application. Besides, the application of multi-level attention in our WSDL approach boosts the localization of discriminative regions and further improves the fine-grained image classification accuracy.

Comparing with the methods using ResNet [64], our WSDL approach still achieves better performance. RBF proposes a non-parametric method for metric learning and classification, which is based on the ResNet-50. But it is 6.73% lower than our WSDL approach. AGAL also uses ResNet-50, and achieves adjacent classification accuracy with ours (85.40% and 85.50%). But, it is noted that part locations and attribute annotations are used in AGAL, neither of them is used in our WSDL approach.

The experimental results of comparisons with state-of-the-art methods on Cars-196 dataset in the aspect of classification accuracy are shown in Table II. The trends are similar with CUB-200-2011 dataset, where our WSDL approach achieves the best classification accuracy among all the state-of-the-art methods under the same weakly supervised setting, which brings 1.00% improvement than the best classification results from compared methods. Our WSDL approach outperforms those methods using object annotations, such as DPL-CNN [47], [61], and is only beaten by PG Alignment [8] and BoT [60] no more than 0.30%.

From Tables I and II, we can see that our WSDL approach with n-pathway achieves better classification accuracy than our conference version [16] on both CUB-200-2011 and Cars-196 datasets, which verifies the effectiveness of n-pathway approach with multi-level attentions.

*2) Efficiency of Classification:* Experimental results of comparisons with state-of-the-art methods in the aspect of classification efficiency are presented in Table III. Average classification speed is evaluated by the frames recognized per second, denoted as fps. Since it has little relation with datasets, CUB-200-2011 dataset is adopted as the evaluation dataset. We get the average classification speed on the computer with one GPU of NVIDIA TITAN X @1417MHZ and one CPU of Intel Core i7-6900K @3.2GHZ. Compared with state-of-the-art methods, our WSDL approach achieves the best performance on not only the classification accuracy but also the efficiency. We split state-of-the-art methods into 2 groups by the basic CNN models used in their methods: VGGNet [39] and AlexNet [65]. Apart from hardware environment, average classification speed also depends on implementation of the method. Different implementations achieve different average classification speeds. For fair comparison, we directly run the source codes provided by authors of compared methods under the same experimental setting, except Part-stacked CNN [55]. Its average classification speed is reported as 20 fps in the original paper. It reports that a single CaffeNet [66] runs at 50 fps under the experimental setting (NVIDIA Tesla K80). In our experiments, a single CaffeNet runs at 35.75 fps, so we calculate the speed of Part-stacked CNN in the same experimental setting with ours as $20 \times 35.75 \div 50 = 14.30$ fps. Compared with state-of-the-art methods in the first group, our WSDL approach is 2 × faster than Bilinear-CNN (9.09 fps vs. 4.52 fps). Besides, the classification accuracy of our WSDL approach is also 1.61% higher than Bilinear-CNN on CUB-200-2011 dataset. Even more, our WSDL approach is over two orders of magnitude faster than methods based on labor-consuming region proposal methods, such as TSC [12], TL Atten [9] and NAC [11]. When utilizing AlexNet as the basic CNN, our WSDL approach is still faster than Part-stacked CNN [55], which also utilizes AlexNet. It is noted that neither object nor part annotations are used in our approach, while all are used in Part-stacked CNN. Our WSDL approach avoids the time-consuming classification process by the design of discriminative localization network (DLN) with one region proposal network and multiple localization networks, and achieves the best classification accuracy by the mutual promotion between localization and classification. This leads the fine-grained image classification to practical application.

TABLE IV

EFFECTIVENESS OF MULTI-LEVEL ATTENTION IN OUR WSDL APPROACH ON CUB-200-2011 AND CARS-196 DATASETS IN THE ASPECT OF CLASSIFICATION ACCURACY

| Convolutional layers | Accuracy (%) | |
|---|---|---|
| | CUB-200-2011 | Cars-196 |
| Conv_cam | 83.45 | 89.59 |
| Conv5_3 | 81.15 | 84.31 |
| Conv4_3 | 77.84 | 78.01 |
| Conv_cam + Conv5_3 | 84.43 | 90.29 |
| Conv_cam + Conv4_3 | 84.36 | 90.10 |
| Conv4_3 + Conv5_3 | 81.41 | 84.68 |
| Conv_cam + Conv5_3 + Conv4_3 | 84.59 | 90.30 |

TABLE V

EFFECTIVENESS OF DISCRIMINATIVE LOCALIZATION NETWORK IN OUR WSDL APPROACH ON CUB-200-2011 DATASET IN THE ASPECT OF CLASSIFICATION EFFICIENCY

| Methods | Average Classification Speed (fps) |
|---|---|
| one-level | 10.07 |
| two-level (respectively) | 5.04 |
| two-level (with dln) | 9.09 |
| three-level (respectively) | 3.36 |
| three-level (with dln) | 7.69 |

### *D. Effectivenesses of Components in Our WSDL Approach*

Detailed experiments are performed to show the effectiveness of each component in our WSDL approach in the following two aspects:

*1) Effectiveness of Multi-Level Attention in the Aspect of Classification Accuracy:* In our WSDL approach, multi-level attention is applied. Different level attentions focus on different characteristics of the image, which are complementary and boost the classification accuracy. In the experiments, we extract the attention maps from the convolutional layers of "Conv4_3", "Conv5_3" and "Conv_cam" in our MAEN, and evaluate their effectivenesses. From Table IV, we can observe that the combination of different level attentions boosts the classification accuracy, which verifies the complementarity among them. The attention from "Conv4_3" plays a minor role in promoting the classification accuracy. Besides, the time consumption of the application of three-level attention is high. Therefore, in our experiments, we only adopt two-level attention from "Conv5_3" and "Conv_cam" to achieve the best trade-off between classification accuracy and efficiency, as shown in Tables I to III.

*2) Effectiveness of Discriminative Localization Network in the Aspect of Classification Efficiency:* Since we apply multi-level attention, there are 2 choices: (1) Train $n$ discriminative localization networks, each of which consists of one RPN and one Fast R-CNN, denoted as "two-level (respectively)" and "three-level (respectively)" in Table V, which causes the linear growth of time consumption. (2) In our WSDL approach, we design an $n$-pathway discriminative localization network with one RPN and $n$ localization networks, and all of them share the same region proposals generated by RPN, which avoids the linear growth of time consumption, denoted as

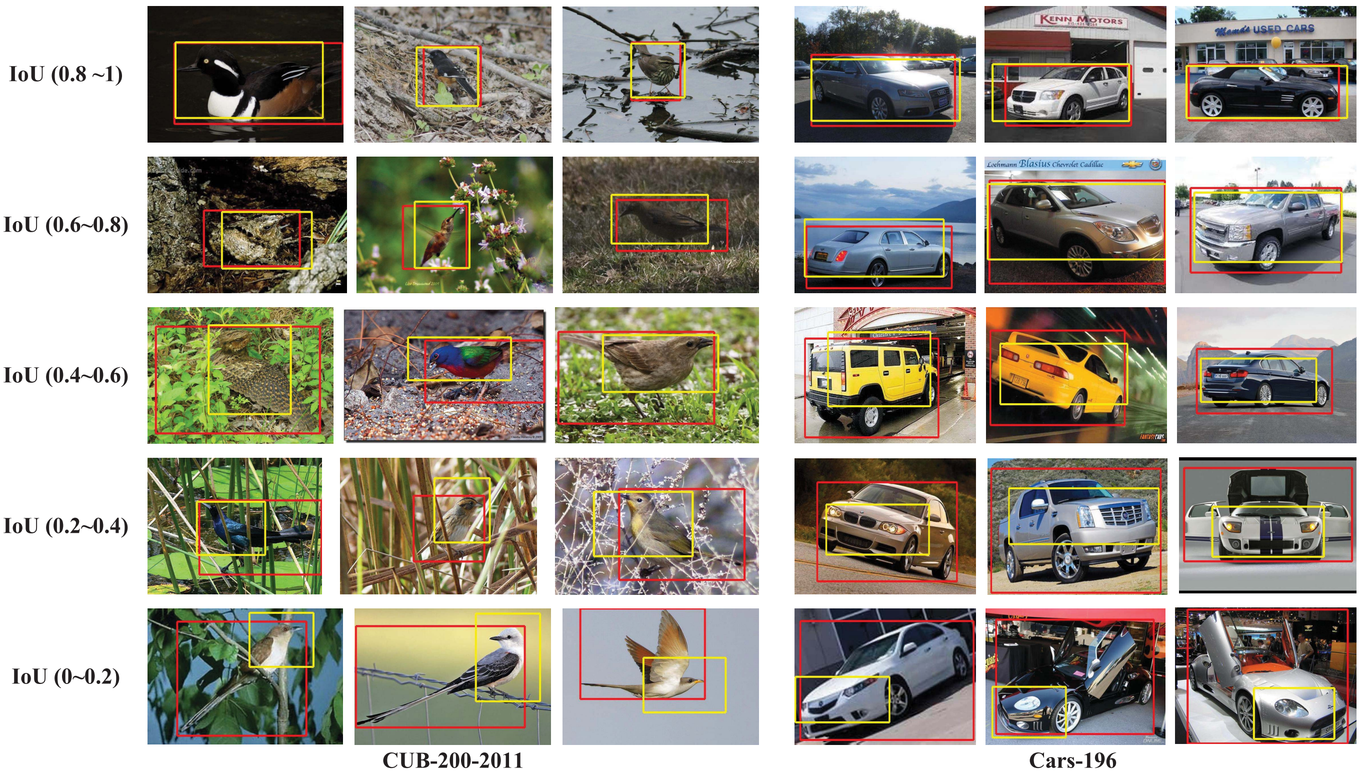

Fig. 4. Samples of predicted bounding boxes of discriminative regions (yellow rectangles) based on the attention information from "Conv_cam" and ground-truth bounding boxes of objects (red rectangles) at different ranges of IoU on CUB-200-2011 and Cars-196 datasets.

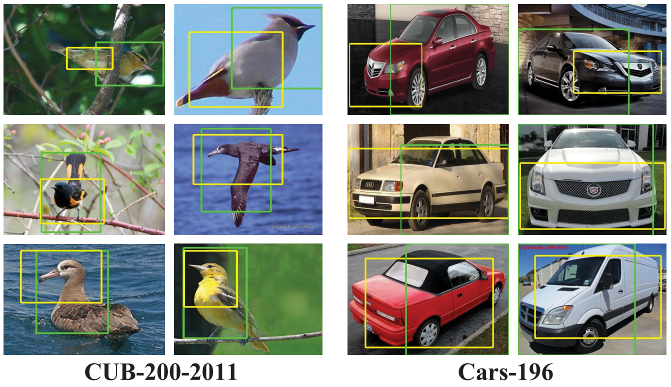

Fig. 5. Samples of the bounding boxes of discriminative regions based on two level attentions from "Conv_cam" (yellow rectangles) and "Conv5_3" (green rectangles) on CUB-200-2011 and Cars-196 datasets.

"two-level (dln)" and "three-level (dln)" in Table V. From Table V, we can observe that our designed architecture of DLN reduces the time consumption.

### E. Comparisons With Baselines

Our WSDL approach is based on Faster R-CNN [26], MAEN, and VGGNet [39]. To verify the effectiveness of our WSDL approach, we present the results of comparisons with Faster R-CNN, MAEN and VGGNet on CUB-200-2011 dataset in Table VII. "VGGNet" denotes the result of directly using fine-tuned VGGNet, "MAEN" denotes the result of directly using MAEN, and "Faster R-CNN (gt)" denotes the result of directly adopting Faster R-CNN with ground-truth bounding box of the object to guide training phase. Our WSDL approach achieves the best performance even without using object or part annotations. We adopt VGGNet as the basic model in our approach, but its classification accuracy is only 70.42% , which is much lower than ours. It shows that the discriminative localization has promoting effect to classification. With discriminative localization, we find the most important regions of images for classification, which contain the key variance from other subcategories. Compared with "Faster R-CNN (gt)", our approach also achieves better performance. It is an interesting phenomenon that worth thinking about. From the last row in Fig. 4, we observe that not all the areas in the ground-truth bounding boxes are necessary for classification. Some ground-truth bounding boxes contain large area of background noise that has less useful information and even leads to misclassification. So discriminative localization is significantly helpful for achieving better classification performance. MAEN has similar localization accuracy with our WSDL approach according to Table VIII, but has lower accuracy as in Table VII. It is mainly because of the different learning abilities of MAEN and WSDL. In the training phase, MAEN only learn from the original images. While our WSDL approach first generates proposals for each image, and these proposals drive the model to learn discriminative localization and classification simultaneously, which makes the model learn more discriminative features with multiple scales and granularities. So our WSDL approach can achieve better classification accuracy.

TABLE VI

PCL FOR EACH PART OF THE OBJECT IN THE CUB-200-2011 DATASET

| Parts | back | beak | belly | breast | crown | forehead | left eye | left leg |
|---|---|---|---|---|---|---|---|---|
| PCL (%) | 96.33 | 96.49 | 94.00 | 95.29 | 97.38 | 97.07 | 97.49 | 89.92 |
| Parts | left wing | nape | right eye | right leg | right wing | tail | throat | **average** |
| PCL (%) | 92.60 | 96.60 | 96.79 | 91.85 | 97.00 | 85.03 | 96.38 | **94.68** |

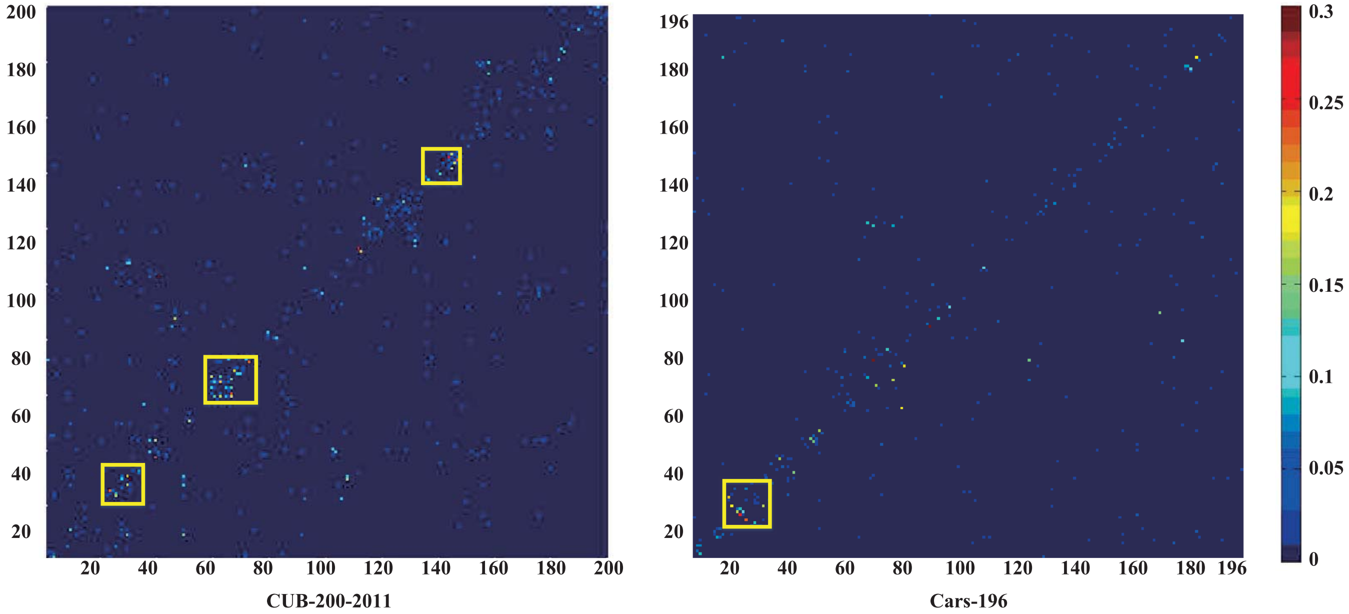

Fig. 6. Classification confusion matrices on CUB-200-2011 and Cars-196 datasets. The yellow rectangles show the sets of subcategories with the higher probability of misclassification.

### F. Effectiveness of Discriminative Localization

Our WSDL approach focuses on improving the localization and classification performance simultaneously. Since the discriminative regions are generally located at the region of the object in the image, we adopt the IoU overlap between the discriminative region and ground-truth bounding box of the object to evaluate the correctness of localization. We consider a bounding box of discriminative region to be correctly predicted if its IoU with ground-truth bounding box of the object is larger than 0.5. We show the results obtained from "Conv_cam" on CUB-200-2011 and Cars-196 datasets in Table VIII, and our WSDL approach achieves the accuracy of 46.05% and 56.60%. Considering that neither object nor part annotations are used, it is a promising result. Compared with "MAEN" which means directly using the attention map from "Conv_cam" to generate bounding box, our WSDL approach achieves improvements by 8.21%, which verifies its effectiveness.

We also show some samples of predicted bounding boxes of discriminative regions and ground-truth bounding boxes of objects at different ranges of IoU (e.g. 0∼0.2, 0.2∼0.4, 0.4∼0.6, 0.6∼0.8, 0.8∼1) on CUB-200-2011 and Cars-196 datasets, as shown in Fig. 4. We have some predicted bounding boxes whose IoUs with ground-truth bounding boxes of objects are lower than 0.5. But these predicted bounding boxes contain discriminative regions of the objects, such as heads or bodies. It verifies the effectiveness of our WSDL approach in localizing discriminative regions of object for achieving better classification performance. Fig. 5 shows the bounding boxes of discriminative regions based on two level attentions from "Conv_cam" and "Conv5_3". We can observe that different attentions focus on different regions, and provide complementary information to boost the classification accuracy. To further verify the effectiveness of discriminative localization in our WSDL approach, quantitative results are given in terms of the Percentage of Correctly Localization (PCL) in Table VI, which estimates whether the predicted bounding box contains the parts of object or not. CUB-200-2011 dataset provides 15 part locations, which denote the pixel locations of centers of parts. We consider our predicted bounding box contains a part if the part location lies in the area of the predicted bounding box. Table VI shows that about average 94.68% of the parts located in our predicted bounding boxes. It shows that our discriminative localization can detect the distinguishing information of objects to promote classification performance.

### G. Different Focuses of Different Level Attentions

As described in [32] and [67], different convolutional layers capture patterns from simple visual elements such as edges, to complex visual concepts such as parts and objects. Different layers describe the visual content at different parts, each of which is complementary to each other for the task of recognition. We generate bounding boxes from different convolutional

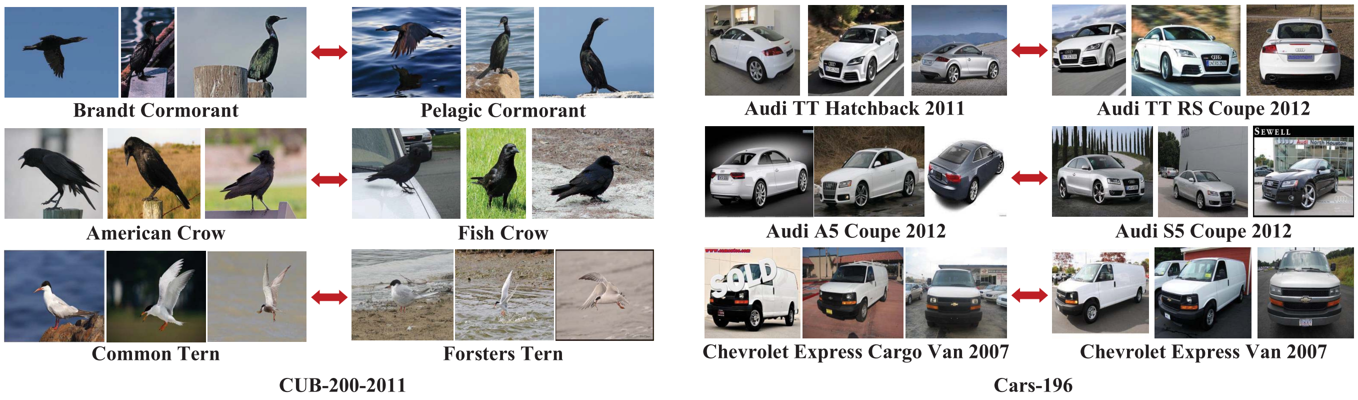

Fig. 7. Examples of the most confused subcategory pairs in CUB-200-2011 and Cars-196 datasets. One subcategory is mostly confidently classified as the other in the same row in the testing phase.

TABLE VII
COMPARISON WITH BASELINES ON CUB-200-2011 DATASET

| Methods | Accuracy (%) |
|---|---|
| **Our WSDL Approach** | **85.71** |
| Faster R-CNN (gt) | 82.46 |
| MAEN | 77.50 |
| VGGNet | 70.42 |

TABLE VIII
LOCALIZATION RESULTS ON CUB-200-2011 AND CARS-196 DATASETS

| Methods | Localization Accuracy (%) | |
|---|---|---|
| | CUB-200-2011 | Cars-196 |
| **Our WSDL Approach** | **46.05** | **56.60** |
| MAEN | 44.93 | 55.79 |

TABLE IX
PROPORTIONS OF IOU ON CUB-200-2011 AND CARS-196 DATASETS

| IoU<br>Dataset | >0.5 | >0.6 | >0.7 | >0.8 | >0.9 |
|---|---|---|---|---|---|
| CUB-200-2011 | 13.22% | 6.13% | 2.28% | 0.50% | 0 |
| Cars-196 | 4.44% | 0.85% | 0 | 0 | 0 |

layers, which have different focuses on different regions. We adopt Intersection-over-Union (IoU) [37] of bounding boxes from the layer of "Conv_cam" and "Conv5_3", to verify different focuses of different level attentions. In Table IX, we present the proportion when IoU is over a specific threshold. We can see that the proportions of $IoU > 0.5$ are only 13.22% in CUB-200-2011 dataset and 4.44% in Cars-196 dataset. The proportions of $IoU > 0.7$ are very small, which verify that different level of activation maps attend to different discriminative regions.

There may be more than one connected areas and they are both discriminative regions. They will be obtained by different convolutional layers. Actually, we consider one discriminative region in each convolutional layer, which has the largest connected area and would be the most discriminative region. But, different convolutional layers focus on different regions as described above. Therefore, multi-level attention could cover the discriminative regions.

### *H. Analysis of Misclassification*

Fig. 6 shows the classification confusion matrices on CUB-200-2011 and Cars-196 datasets, where coordinate axes denote subcategories and different colors denote different probabilities of misclassification. The similar subcategories that belong to the same genus or car brand are set to the adjacent image-level subcategory label ids in the original datasets. So misclassified subcategories with higher probability would appear near the diagonal, as shown in the yellow rectangles of Fig. 6. The small variance is not easy to measure from the image, which leads the high challenge of fine-grained image classification. Fig. 7 shows some examples of the most probably confused subcategory pairs. One subcategory is most confidently classified as the other in the same row. The subcategories in the same row look almost the same, and belong to the same genus. For example, "Common Tern" and "Forsters Tern" look the same in the appearance, as shown in the left third row of Fig. 7, because both of them have the same attributes of white wings and black forehead, and belong to the genus of "Tern". It is even extremely difficult for human beings to distinguish between them. Similarly, it is hard to distinguish between "Audi TT Hatchback 2011" and "Audi TT RS Coupe 2012".

## V. CONCLUSION

In this paper, the weakly supervised discriminative localization approach (WSDL) has been proposed for fast fine-grained image classification. We first apply multi-level attention to guide the discriminative localization learning to localize multiple discriminative regions simultaneously for each image, which only uses image-level subcategory label to avoid using labor-consuming annotations. Then we design an $n$-pathway end-to-end discriminative localization network to simultaneously localize discriminative regions and encode discriminative features, which not only achieves a notable classification performance but also improves classification speed. Their combination simultaneously improves classification speed and eliminates dependence on object and part annotations. Comprehensive experimental results show our WSDL approach is more effective and efficient compared with state-of-the-art methods on 2 widely-used datasets.

The future works lie in three aspects: First, we will make our WSDL approach better by learning better discriminative localization via exploiting the effectiveness of fully convolutional networks. Second, we will make our WSDL approach faster by designing a more efficient network with less operations for a forward pass. Third, we will make our WSDL approach more generalized by training one model to support the classification of different datasets.

## 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] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Sep. 1999, pp. 1150–1157.

[3] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[4] N. Sudha, A. R. Mohan, and P. K. Meher, "A self-configurable systolic architecture for face recognition system based on principal component neural network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 8, pp. 1071–1084, Aug. 2011.

[5] W. Paier, M. Kettern, A. Hilsmann, and P. Eisert, "A hybrid approach for facial performance analysis and editing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 4, pp. 784–797, Apr. 2017.

[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.*, 2014, pp. 834–849.

[7] X. Zhang, H. Xiong, W. Zhou, and Q. Tian, "Fused one-vs-all mid-level features for fine-grained visual categorization," in *Proc. ACM Int. Conf. Multimedia (ACM MM)*, 2014, pp. 287–296.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] X. He and Y. Peng, "Weakly supervised learning of part selection model with spatial constraints for fine-grained image classification," in *Proc. Conf. Artif. Intell. (AAAI)*, 2017, pp. 4075–4081.

[13] J. R. R. Uijlings, K. E. A. 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, Apr. 2013.

[14] 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.

[15] 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.

[16] X. He, Y. Peng, and J. Zhao, "Fine-grained discriminative localization via saliency-guided faster R-CNN," in *Proc. ACM Multimedia Conf. (ACM MM)*, 2017, pp. 627–635.

[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] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[20] 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.

[21] T. Berg and P. 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.

[22] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1. Jun. 2005, pp. 886–893.

[23] 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.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[25] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 391–405.

[26] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 91–99.

[27] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free?—Weakly-supervised learning with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 685–694.

[28] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1717–1724.

[29] 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.

[30] S. Zagoruyko and N. Komodakis, "Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–13.

[31] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[32] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4305–4314.

[33] J. R. Anderson, *Cognitive Psychology and Its Implications*. San Francisco, CA, USA: Freeman, 1990.

[34] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *Artif. Intell.*, vol. 78, nos. 1–2, pp. 507–545, Oct. 1995.

[35] Y. Karklin and M. S. Lewicki, "Emergence of complex cell properties by learning to generalize in natural scenes," *Nature*, vol. 457, no. 7225, pp. 83–86, 2009.

[36] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

[37] M. Everingham, S. A. Eslami, L. V. Gool, C. K. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, 2015.

[38] 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.

[39] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[40] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3D object representations for fine-grained categorization," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 554–561.

[41] 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.

[42] M. Jaderberg *et al.*, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2015, pp. 2017–2025.

[43] 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.

[44] 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.

[45] B. J. Meyer, B. Harwood, and T. Drummond. (2017). "Nearest neighbour radial basis function solvers for deep neural networks." [Online]. Available: https://arxiv.org/abs/1705.09780

[46] 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.

[47] 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.

[48] L. Xie, L. Zheng, J. Wang, A. Yuille, and Q. Tian, "Interactive: Inter-layer activeness propagation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 270–279.

[49] 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.

[50] Y. Cui, F. Zhou, Y. Lin, and S. Belongie. (2015). "Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop." [Online]. Available: https://arxiv.org/abs/1512.05227

[51] X. Liu, J. Wang, S. Wen, E. Ding, and Y. Lin, "Localizing by describing: Attribute-guided attention localization for fine-grained recognition," in *Proc. AAAI*, 2017, pp. 4190–4196.

[52] 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.*, vol. 40, no. 5, pp. 1100–1113, May 2018.

[53] S. Branson, G. Van Horn, S. Belongie, and P. Perona. (2014). "Bird species categorization using pose normalized deep convolutional nets." [Online]. Available: https://arxiv.org/abs/1406.2952

[54] 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.

[55] 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.

[56] B. Zhao, X. Wu, J. Feng, Q. Peng, and S. Yan. (2016). "Diversified visual attention networks for fine-grained object classification." [Online]. Available: https://arxiv.org/abs/1606.08572

[57] 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.

[58] Q. Hu, H. Wang, T. Li, and C. Shen, "Deep CNNs with spatially weighted pooling for fine-grained car recognition," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 11, pp. 3147–3156, Nov. 2017.

[59] J. Krause, T. Gebru, J. Deng, L.-J. Li, and L. Fei-Fei, "Learning features and parts for fine-grained recognition," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2014, pp. 26–33.

[60] 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.

[61] Y. Wang, V. I. Morariu, and L. S. Davis. (2016). "Weakly-supervised discriminative patch learning via CNN for fine-grained recognition." [Online]. Available: https://arxiv.org/abs/1611.09932

[62] 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.

[63] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. (2014). "Return of the devil in the details: Delving deep into convolutional nets." [Online]. Available: https://arxiv.org/abs/1405.3531

[64] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[65] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2012, pp. 1097–1105.

[66] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia (ACM MM)*, 2014, pp. 675–678.

[67] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 818–833.

**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.

**Yuxin Peng** received the Ph.D. degree in computer application from Peking University, Beijing, China, in 2003. After that, he was an Assistant Professor with the Institute of Computer Science and Technology (ICST), Peking University, where he was promoted to an Associate Professor and a Professor in 2005 and 2010, respectively. He is currently a Professor with ICST, Peking University, and a Chief Scientist in the 863 Program (National Hi-Tech Research and Development Program of China). 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. He has published over 100 papers in refereed international journals and conference proceedings, including IJCV, TIP, TMM, TCSVT, PR, ACM MM, ICCV, CVPR, IJCAI, AAAI, and so on. He has submitted 35 patent applications and received 16 of them. His current research interests mainly include cross-media analysis and reasoning, image and video analysis and retrieval, and computer vision. He led his team to participate in TREC Video Retrieval Evaluation (TRECVID) many times. In TRECVID 2009, his team received four first places on four sub-tasks of the High-Level Feature Extraction task and the Search task. In TRECVID 2012, his team received four first places on all four sub-tasks of the Instance Search (INS) task and the Known-Item Search task. In TRECVID 2014, his team received the first place in the Interactive INS task. His team also received both first places in the INS task of TRECVID 2015, 2016, and 2017. In addition, he received the first prize of the Beijing Science and Technology Award in 2016 (ranking first).

**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.