Task-Driven Progressive Part Localization for Fine-Grained Object Recognition

Chen Huang, *Student Member, IEEE,* Zhihai He, *Fellow, IEEE* and Wenming Cao, *Member, IEEE,*

***Abstract*—The problem of fine-grained object recognition is very challenging due to the subtle visual differences between different object categories. In this paper we propose a task-driven progressive part localization (TPPL) approach for fine-grained object recognition. Most existing methods follow a two-step approach which first detects salient object parts to suppress the interference from background scenes and then classify objects based on features extracted from these regions. The part detector and object classifier are often independently designed and trained. In this paper, our major finding is that the part detector should be jointly designed and progressively refined with the object classifier so that the detected regions can provide the most distinctive features for final object recognition. Specifically, we develop a part-based SPP-net (Part-SPP) as our baseline part detector. We then establish a task-driven progressive part localization framework, which takes the predicted boxes of Part-SPP as an initial guess, then examines new regions in the neighborhood using a particle swarm optimization approach, searching for more discriminative image regions to maximize the objective function and the recognition performance. This procedure is performed in an iterative manner to progressively improve the joint part detection and object classification performance. Experimental results on the Caltech-UCSD-200-2011 dataset demonstrate that our method outperforms state-of-the-art fine-grained categorization methods both in part localization and classification, even without requiring a bounding box during testing.**

***Index Terms*—fine-grained recognition, deep learning, regional convolutional neural network, spatial pyramid pooling, deformable part-based model**

I. Introduction

Fine-grained object categorization, also refered to as subcategory recognition, is a rapidly growing subfield in multimedia content analysis. Different from traditional image classification such as scene or object recognition, fine-grained categorization deals with images which have subtle distinctions, such as recognizing bird species [1]–[6], the breed of a dog or cat [7]–[9], or the model of an aircraft [10], [11]. Fine-grained object recognition has important applications in practice [3], [9], [12]. Even in the ImageNet Challenge, an important issue faced by state-of-the-art recognition algorithms is how to distinguish closely related subcategories [13]. Fine-grained categorization is a challenging computer vision problem since it needs to deal with subtle differences in the overall appearance between classes (small inter-class variations) and large appearance variations within the same class (large intra-class variations). Therefore, it requires methods that are more discriminative than *basic-level* image classification to classify differences between highly similar subcategories. Moreover, these differences are often overwhelmed by nuisance factors such as variations in poses, viewpoints or locations of objects in the images [14].

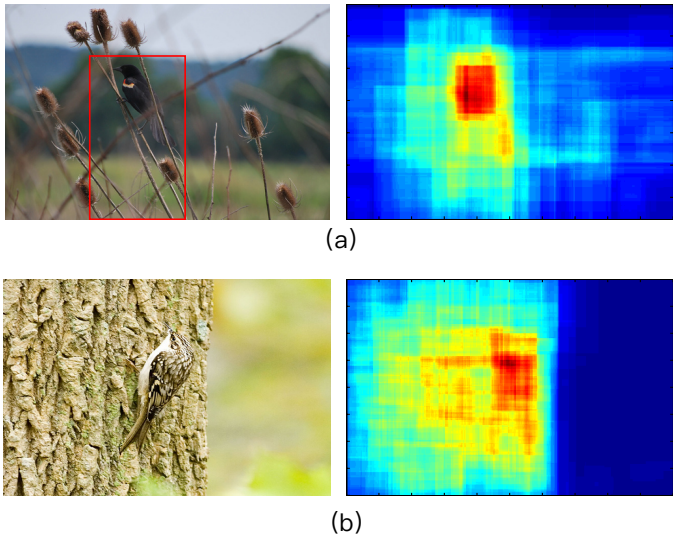


Fig. 1. Samples to show the correlation between object part detection and classification. The left images are detection results. In these cases detection are failed, since the target is either too small, or is too visually similar to the background. The images on the right are classification score maps for the correspondence category. They show which region contributes most to the categorization decision. Our intuition here is that we need to utilize such information to correct or refine the detection results.

Most visual object recognition methods are mainly based on global representation containing statistics of local features calculated in the whole image [15]–[19]. However, for fine-grained object categorization, this approach is not effective, since global appearances are highly similar for different subcategories and only small differences at certain locations allow for discrimination. Therefore, using precise part information has become an important approach for fine-grained object recognition, especially for handling the issues of pose and viewpoint variations. A common procedure [1], [20]–[22] is to first localize various semantic parts and then model the appearance variation conditioned on their detected locations. Recently, part-based approaches and their variants [23]–[25] have demonstrated significantly performance improvement

C. Huang and Z. He are with the Department of Electrical and Computer Engineering, University of Missouri, Columbia, MO, 65211, USA, e-mail: hezhi@missouri.edu

W. Cao are with the College of Information Engineering, Shenzhen University, P. R. China, e-mail: caom@shenzhenu.edu.cn

over earlier work [26], [27].

It should be noted that, in existing part-based object recognition approaches, the part detectors which localize the semantic parts, such as bird heads and bodies, and the object recognition which classifies the objects based on features extracted from the part regions, are often designed independently. Certainly, these two tasks are different, however, they are closely related. If one has a good object detector, it becomes easy to predict image labels when objects are accurately located with high scores. Inversely, knowing the class of an image can help to detect object parts. Fig. 1 shows two examples where the detection tasks failed; however, classification score response maps include a strong clue on where the object is located. Moreover, we recognize that in some cases even the detected parts are semantically accurate, it may not be optimal for classification, as these parts are subjectively defined by human labeling, and the part detection module is trained from these hand-labeled samples. In this work, we note that the feedback information from the classifier could be utilized to correct or refine these inaccurate or inefficient part detection results.

Based on this observatoin, we propose to explore a task-driven approach for progressive part detection where the task of the part detector is to detect image regions that provide the most discriminative visual features for the subsequent object classification. In other words, the detected parts may not match the ground-truth parts specified by human. However, these image regions are most useful for object classification.

Specifically, in this paper, we investigate how to best localize discriminative parts for fine-grained image categorization. We propose a recognition task-driven progressive part localization(TPPL) framework, in which detection and classification are refined jointly and progressively. We start with a revised version of Part-based R-CNN [24] and SPP-net [28] to generate initial detection results. We developed a particle swarm optimization algorithm to refine the part localization results by proposing new part regions with scores that are likely to give better classification performance than the original ones. By doing so, even if the initial detection results are not good, the algorithm can find regions that contain more discriminative parts after a few iterations. After that, deep convolutional neural network (DCNN) features are then extracted from these regions for final object classification. Experimental results on the Caltech-UCSD-200-2011 dataset demonstrate that our method outperforms state-of-the-art fine-grained categorization methods in both part localization and classification. The basic framework of our proposed method is illustrated in Fig. 2.

A preliminary version of this manuscript has been presented in a conference [29]. The major contributions of this paper are: (1) we recognize the inherent correlation between accurate part localization and fine-grained categorization and the need for joint design of these two components. (2) We propose a TPPL algorithm that finds more discriminative part regions. (3) The aforementioned method is complementary and can be easily adopted to various part-based detection approaches. (4) We demonstrate significant improvement in classification performance over state-of-the-art methods on the CUB-2011 benchmark dataset.

The rest of this paper is organized as follows: in Section II, we review the related work on fine-grained object recognition. Section III presents the correlation between accurate part localization and fine-grained categorization; then introduces our TPPL framework. We present experiments and analysis on the CUB-2011 datasets in Section IV and conclude with future work in Section V.

II. Related work

Existing work related to this paper lies in the following three areas: (A) fine-grained recognition with global features; (B) semantic part localization; (C) feature representation of part regions for object classification.

A. Fine-grained Recognition with Global Features. Fine-grained image categorization has been extensively studied recently. Early work in this area concentrated on direct application of feature description and encoding methods for object categorization [3], [12], [15]–[17]. For example, [17] developed a new image feature encoding method based on Fisher kernels which describes image patches by their deviation from an universal generative Gaussian mixture model. Zhou *et al.* developed a super vector approach which maps each image region descriptor into a high-dimensional sparse vector and then aggregate vectors from all regions to form an image-level global feature for image classification [15]. Due to the visual similarity between closely related classes, there is a high likelihood for having a significant amount of common visual words shared between classes that do not help distinguishing these categories from each other. To address this limitation, Yao *et al.* [30] proposed a code-free approach, which applies a template matching process to the whole image with a large number of randomly generated image templates. The approach then uses a bagging-based algorithm to build a final classifier by aggregating a set of discriminative yet largely uncorrelated classifiers. A stacked evidence tree method has been developed in [31] which trains a random forest of trees to predict the class of an image based on individual keypoint descriptors. During image classification, descriptors for all detected keypoints are dropped through trees, and the evidence at each leaf is accumulated to obtain an overall evidence vector for classification.

To minimize the impact of cluttered background on the classification of foreground objects, Chai *et al.* have developed a bi-level co-segmentation algorithm to remove the background [3]. A Hierarchical Part Matching (HPM) method has been developed in [32] to capture better representation of foreground body parts based on part-based image alignment and segmentation. They use a hierarchical structure learning (HSL) algorithm to find mid-level concepts beyond basic parts and a geometric phrase pooling (GPP) algorithm to aggregate mid-level structures in the local feature groups for fine-grained object classification. Recently, it has been observed that these types of segmentation-based methods often suffer from the risk of losing subtle differences between subcategories [33], [34].

B. Semantic Part Localization. Incorporating precise part information into the classification framework has proved to be critical for building accurate fine-grained recognition systems

in recent studies. In their early work, Felzenszwalb *et al.* [35] developed discriminatively trained part models using latent SVM. Bourdev *et al.* developed a part-based approach for characterizing persons [27] which decomposes the image into a set of parts, called poselets, each capturing a salient pattern corresponding to a given viewpoint and local pose. They combine evidences from different parts of the body at different scales, which provides a robust distributed representation of a person. Pose-normalized image description has been introduced in [20] for fine-grained object recognition using deformable part descriptors (DPD) and deformable part matching (DPM). Recently, the Part-based R-CNN [24] has achieved the state of the art performance on part detection using R-CNN [36]. This method first proposes thousands of candidate bounding boxes for each image using some bottom-up region proposal approaches, then selects the bounding boxes with high classification scores as the detection results. In our work, we modify the Part-based R-CNN, by replacing the R-CNN with a much faster SPP-net [28] as the baseline. The SPP-Net uses a spatial pyramid pooling method to handle different input image sizes and increase the robustness under object deformation. Starting with the part localization results of Part-based SPP-net as our initial guess, we refined it using an iterative search to find more discriminative regions for the later recognition task.

C. Part Feature Representation. After part detection, the subsequent step is to construct efficient feature representation of part regions for classification. Berg *et al.* has proposed the part-based one-vs-one features library POOF [1] for the mid-level features. Each of these features specializes in discrimination between two particular classes based on the appearance at a particular part. For any pair of classes and for any pair of parts, they extract low-level features on a grid of cells that covers the two parts, and train a linear classifier between these two classes. The weights assigned by this classifier to different cells of the grid indicate the most discriminative region around these parts for this pair of classes. Recently, deep convolutional neural network (CNN) has become the dominant approach for large-scale image feature extraction and classification since the success of Krizhevsky *et al.* [37] on ImageNet recognition with CNNs. The network structure of Krizhevsky *et al.* consists of five convolutional layers and two fully-connected layers, followed by a softmax layer to predict the class label. Although this network is still popular, there have been efforts to improve the CNN architecture. For example, recent works use more layers [24], [38], [39] to achieve even better performance. In this paper, we use a seven-layer network structure in our experiment, but our approach proposed here is expected to be applicable to CNNs with deeper architectures.

III. Our Approach

In the following, we first introduce our Part-based SPP-net (Part-SPP), a modified version of Part-based R-CNN which uses the SPP-net for faster part detection and better performance. This is our baseline method. We then analyze the underlying assumption of the strong correlation between part localization and fine-grained classification. Motivated by previous analysis, the TPPL framework is proposed to further boost fine-grained classification accuracy by refining detection results and making them optimal for the recognition task. Finally, we combine all parts' classification scores in a discriminative way to make decisions for the final categorization results.

A. Baseline Part Detection Using SPP-net

Part-based RCNN is one of the state-of-the-art approaches for fine-grained object recognition. It uses a part detection model, generalizes the R-CNN framework [36] to detect parts in addition to the whole object, and enforces geometric constraints between different parts. However, the feature computation in R-CNN is very time-consuming, because it repeatedly applies the CNNs to raw pixels in thousands of warped regions per image. Furthermore, R-CNN requires a very large amount of space to cache its intermediate features. To address this issue, we chosen the Spatial Pyramid Pooling (SPP-net) [28], which computes feature maps for entire images only once, and then pools features in arbitrary regions (sub-images) using a spatial pyramid pooling approach to generate fixed-length representations for part detection. He *et al.* [28] claims that SPP-net is 24$\sim$102 times faster than the R-CNN method, while achieving better or comparable detection accuracy on the Pascal VOC 2007.

Our Part-SPP baseline is illustrated in Fig 2 (a). It follows the traditional two-stage pipeline: first detecting the semantic parts of objects, extracting features from each localized regions, then classifying each part independently, and finally combining all part classification scores to obtain the final result. So, basically there are two tasks here: 1) detect parts from background, 2) classify localized parts. In our implementation, a SPP-net [28] is trained for the entire object and every part by treating each as a seperate category, which means the same part of different subcategory is considered to be in the same class. For example, "cardinal head" and "American crow head" are both considered as "head" in the detection phase. During part detector training, we use the selective search [40] to generate candidate regions of interest (ROIs) with high objectness scores. In our case every image can generate around 1500 ROIs. Those ROIs which have more than 50% overlap with the ground-truth bounding box are marked as foreground; otherwise, they are considered as background. We use SVM and the hard-negative mining approach to train the SPP-net detector [28]. During testing, all detectors run independently on the image, and each detection score is converted into a probability value using a sigmoid function. Moreover, the joint configuration of the bounding box and its parts is scored as the product of probabilities. We also apply the δ^{box} constraints proposed by Zhang *et al.* [24] as geometric constraints, which sets zero probability for part detections that do not fall within 10 pixels of the bounding box. After detecting object parts, we extract deep CNN features from these localized regions, train one-vs-all SVM part classifiers for each part, and classify them independently. In the classification phase, parts of different subcategories are considered as different classes. Finally, the part classification scores are discriminatively combined for final class label prediction, which will be discussed in Sec III-E.

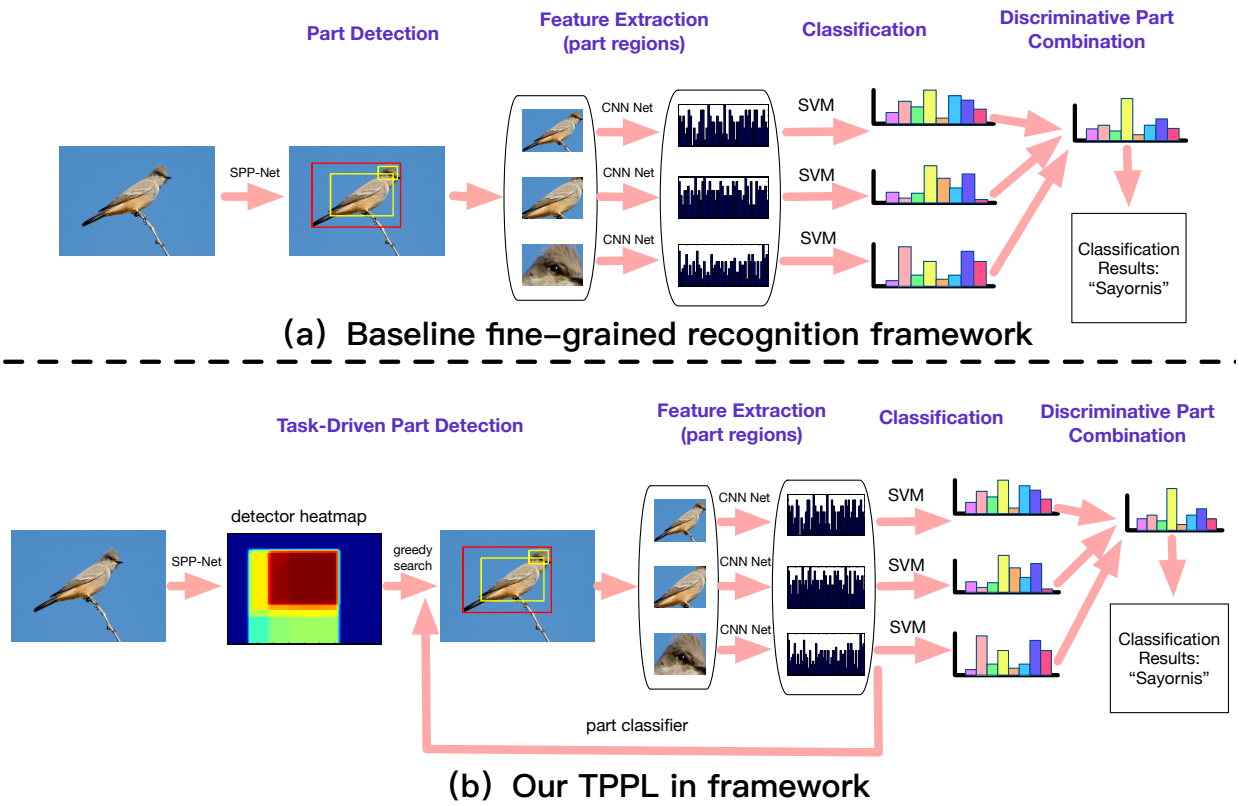


Fig. 2. A conventional part-based recognition pipeline (upper figure) vs our proposed method pipeline (lower figure). In the task-driven part localization framework, part classifier is used to refine the part detector's results. The key for task-driven part localization lie in two folds: 1) how to fuse the results from detectors and classifiers; 2) how to find local maxima. We utilize particle swarm optimization to solve these problems.

Instead of training a network from scratch, we use the ImageNet pre-trained model in our experiments. This architecture is based on the Zeiler and Fergus(ZF) *fast* (smaller) model [41], and the network consists of five convolutional layers. The pooling layer after the last convolutional layer generates 6×6 feature maps, with one 4-level SPP layer, and two 4096-dimension fully connected layers followed by a 1000-way softmax layer. The SPP layer generates a 12,800-dimension representation for each proposed ROI. In order to make the deep CNN-derived features more discriminative for the target task of fine-grained bird classification, we fine-tune the ImageNet pre-trained Zeiler model for the 200-way bird classification task using ground truth bounding box cropping of the Caltech-UCSD Bird 2011 dataset. In particular, we replace the original 1000-way *fc8* classification layer with a new 200-way *fc8* layer with randomly initialized weights drawn from a Gaussian distribution with $\mu = 0$ and $\alpha = 0.01$. We set the fine-tuning learning rate as proposed by SPP-net, initializing the global rate to a tenth of the initial ImageNet learning rate and dropping it by a factor of 10 throughout training, but with a learning rate in the new *fc8* layer 10 times better than that of the global learning rate. For the whole object bounding box and each of the part boxes, we independently fine-tuned the Zeiler model for classification using ground truth bounding box cropping of each region being warped to 224×224, which is the network input size, always with 16 pixels on each edge of the input serving as context in R-CNN. During testing, we extracted features from the detected part regions using that part of the network that had been fine-tuned for that particular body part.

B. Correlation Between Part Detection and Object Classification

It has been well recognized that object part detection is very critical for fine-grained image classification. It is able to suppress the interference from surrounding cluttered background. Previous approaches, which first localized various parts and then modeled the appearance variation conditioned on their detected locations, are highly compatible with this line. All of these methods treat part localization and object classification as two independent tasks. However, we observe that, in many cases both classification and detection can benefit from and contribute to each other's success. This idea relies on the observation that they use different information. This assumption is obvious in fine-grained recognition because detectors are trained to classify foreground and background, while classifiers are learned to distinguish each subcategory object from others, they contain complementary information. Fig. 1 shows two examples where detection tasks fail even though the classification score map for the correspondence category gives a strong clue as to where the objects are. This suggests that it is feasible to use the initially trained object

classifier as an objective function to refine the part detection. The refined part detection with more discrimination power will in turn improve the fine-grained object classification performance. This leads to our task-driven progressive part localization(TPPL) method for object recognition.

Furthermore, in many cases, even the detection results are very accurate, but the later classification task still has a very large probability of failing. In the following, we try to answer two questions: 1) to what extent does accurate detection help classification and 2) how can we refine the current detection algorithm to improve the classification performance?

TABLE I
CORRELATION BETWEEN DETECTION AND CLASSIFICATION FROM TESTS ON THE CUB-200-2011 DATASET.

	Correct Classify	False Classify
Correct Detect	3899	1603
False Detect	153	139

We used the CUB-200-2011 dataset [42] with the SPP-Net baseline method discussed in III-A for part detection. In this experiment, only object bounding boxes were detected as the whole object. We then extracted deep convolutional features inside the predicted bounding boxes and fed them into a one-versus-all linear SVM for classification. The first row of Table I shows that there are 5502 image regions were correctly detected as the body part when compared to the ground truth data provided by human labelling. Within these 5502 image regions, 3899 of them, about 70% were correctly classified while 30% were misclassified. This implies that the current part detection regions do not provide sufficiently accurate representation of the object for effective classification.

The second row of the table shows that there are 292 image regions being incorrectly detected as the body part. However, 48% of them can be still correctly classified during the recognition stage. In [24], even when ground-truth bounding boxes are given, the state-of-the-art object recognition algorithms can only achieve an accuracy of 68.29%. This is mainly because of the following two reasons. (1) Semantic parts are manually defined, so it may not be the most discriminative part of the recognition task. (2) Currently, detection and classification are divided into two separate stages, so even the semantic parts are detected very accurately, it may not be helpful for the subsequent recognition phase. Obviously, if we can develop a scheme to automatically localize the parts which are most discriminative and distinctive for the classification task, we will certainly increase the performance by a large margin. This is the major motivation behind our proposed method.

C. Task-driven Progressive Part Localization

The overall greedy search process of our task-driven progressive part localization is shown in Fig. 3. As in [24], we assume a strongly supervised setting for training, in which we have ground truth bounding box annotations, not only for full objects, but also for a fixed set of semantic parts during the training stage. Given these part annotations, we train one-versus-all linear SVMs for classification. We use the following formula to compute the overall scores for object C_n in the image

$$score(C_n) = \sum_{i=1}^{M} \beta_i \times y(C_n|P_i), \tag{1}$$

where $y(C_n|P_i)$ is the output of the SVM classifier for class C_n for part P_i. M is the number of parts in the image. β_i is the discriminative weight for each part, as we have observed that not all parts of an object were equally useful for recognition. Φ_i is the fine-tuned feature extraction network for each part. so $y(C_n|P_i)$ can be further written as

$$y(C_n|P_i) = \Phi_i(I, P_i) \cdot w_{i,n}. \tag{2}$$

In this equation, $\Phi_i(I, P_i)$ is the deep feature of part P_i on test image I, while $w_{i,n}$ is the trained SVM weight for class C_n on part P_i.

During test, our task is to localize the most distinctive parts for the recognition task. A large number of regions of interest (ROIs) $\{R_1, R_2, ..., R_K\}$ are sampled on the test image using the so-called selective search method developed in [40]. The task of joint detection and recognition then becomes selecting M semantic parts out of K candidate ROIs. However, finding the most distinctive ROIs for a recognition task in hundreds of thousands of candidate regions is extremely expensive in computation since there are $\binom{K}{M}$ possible choices. Thanks to the excellent part localization ability of our baseline Part-SPP, we can only consider ROIs which have a large overlap with initial detection results. By eliminating these low possibility ROIs, our searching range is largely reduced.

Specifically, for each part P_i, the Part-SPP detector will produce a predicted box, shown in the leftmost image of Fig 3. We set this predictive bounding box as an initial guess, and selected candidate regions that have $\geq$ 0.5 overlap with it, donated as $\{T_1, T_2, ...T_J\}$, where J is the number of candidate regions and changes case by case. The second image of Fig 3 shows the current guess (in red) and candidate regions T_j (in blue). The score of each T_j is given by its classification score $y(C_n|T_j) = \Phi_i(I, T_j) \times w_{i,n}$. The region T_j, which has the maximum score, is selected as the current guess for the next iteration. Finally the regions which contain the most distinctive information for recognition are selected, as illustrated in the rightmost image of Fig 3. Note that the detection result is obtained for the recognition task, so it tends to contain as much distinctive parts as possible, rather than trying to have a good overlap with the ground truth bounding box.

For every class C_n, we perform the same task-driven part localization procedure in parallel since the distinctive regions for different classes should be separated from each other. So each class C_n has a classification score

$$score(C_n) = \sum_{i=1}^{M} \beta_i \times y(C_n|P_i) \tag{3}$$

with M supporting regions. By selecting the maximum value in $\arg\max_n score(C_n)$, its index will become the predicted class label for this test image, while its supporting regions will become our part localization result. In this way, the distinctive part localization and object classification are performed jointly. The proposed method is summarized in Algorithm 1.

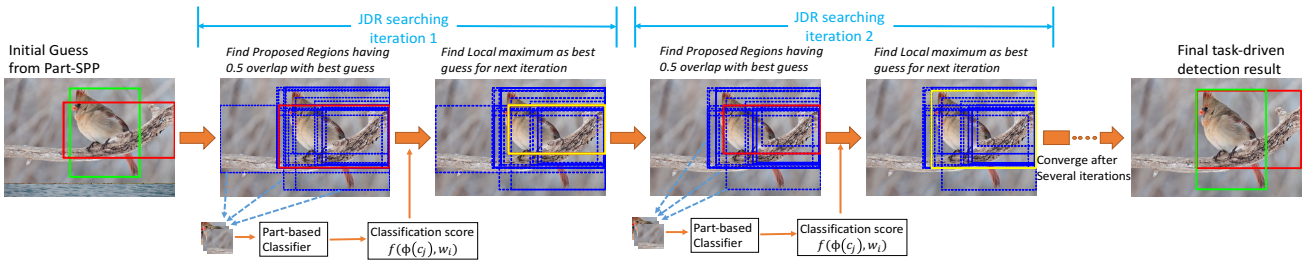


Fig. 3. The overview of our task-driven part progressive localization framework. 1) The detection result of part-SPP is considered an initial guess, shown as a red rectangle in the leftmost image. The green rectangle is the ground-truth bounding box. 2) We selected the proposed regions which have ≥ 0.5 overlap with initial guess as the candidate regions of interest (ROIs), extract deep features $\Phi_i(I, P_j)$ and rank classification scores $y(C_n|P_i) = \Phi_i(I, P_i) \times w_{i,n}$ given by a part-based classifier. 3) The local maxima was chosen as best guess for the next iteration until convergence. The final task-driven detection result is shown as the red rectangle in the rightmost image. Note here this detection result is obtained for recognition task, so it trends to contain as much distinctive part as possible, rather than trying to have good overlap with ground truth bounding box.

D. Particle Swarm Optimization for Progressive Search of the Most Discriminative Part Regions

To ensure high part detection rates, the algorithm often needs to examine a large number of candiate part regions at different locations and sizes generated by the region proposal method [40]. To perform more effective search of the most discriminative part regions for object classification from the large set of candidate regions, we propose to explore the particle swarm optimization approach. Particle swarm optimization (PSO), developed by Kennedy and Eberhart, is an efficient population-based optimization technique which models the set of potential problem solutions as a swarm of particles moving around in a virtual search space. We choose the PSO approach for progressive part localization because: (1) unlike many other algorithms, such as gradient search or convex optimization, the PSO optimization procedure does not require the actual expression of the objective function. Therefore, it can be applied to solve generic nonlinear optimization problems. For example, in this work, we use the classifier as the objective function, which does not have explicit expressions. (2) Research results demonstrate that PSO outperforms other nonlinear evolutionary optimization techniques, such as genetic algorithm and simulated annealing. The PSO is able to handle constraints, especially nonlinear constraints more efficiently than other optimization algorithm by regulating the moves of particles. In part localization, we can include geometric constraints on the body parts. For example, the head part bounding box should be significantly smaller than the body part bounding box. More advanced geometric constraints on the body part regions, such as those used in the part-based R-CNN [24], can be naturally incorporated into PSO to control the moves of particles. (3) The PSO, due to its population-based approach, can avoid local optimum effectively. This is very important for our part localization. As we can see from the heat maps shown in Fig. 3, there are a number of local maximums that could easily trap the progressive part localization, especially in cluttered scenes.

The high-level idea of PSO can be summarized as follows. As illustrated in Fig. 4(a), to find the minimum (or maximum) of an objective function $f(\mathbf{x})$ ($\mathbf{x}$ is a vector) within a solution space, the PSO algorithm starts with a set of candidate solutions called *particles*, $\{\mathbf{x}_p|1 \leq p \leq P\}$ distributed in the solution space. During the optimization process, each particle x_p moves within the solution space in search for the minimum of $f(\mathbf{x})$, and the corresponding movement path is denoted by $x_p(t)$, where t represents time. At each time step, the movement of particle $\mathbf{x}_p$ is given by

$$\mathbf{x}_p(t+1) = \mathbf{x}_p(t) + \mathbf{v}, \tag{4}$$

where

$$\mathbf{v} = w \cdot \mathbf{v} + c_1\Theta_1[\mathbf{x}_p^s - \mathbf{x}_p(t)] + c_2\Theta_2[\mathbf{x}^g - \mathbf{x}_p(t)]. \tag{5}$$

Here, w, c_1 and c_2 are weighting factors, Θ_1 and Θ_2 are two random numbers, $\mathbf{x}_p^s$ is the best solution that the particle itself has found so far, and $\mathbf{x}^g$ is the best solution that all particles have found so far, as illustrated in Fig. 4(a). Each particle, when determining its next move in search for the global optimum, always balances the behaviors of its own and the group.

In this work, during progressive part localization, each particle $\mathbf{x}_p$ represents an image region or a bounding box at location $[x_p, y_p]$ with size $[W_p, H_p]$. Specifically, $\mathbf{x}_p = [x_p, y_p, W_p, H_p]$. The solution space is set to be the larget set of candiate regions generated by the region proposal method, denoted by $\mathbf{S}$. As illustrated in Fig. 4(c), our objective is to find the most discriminative region, or the group best $\mathbf{x}_g$ on the heat map for the test image in (b). We set the number of particles to be 10. During PSO, after we have computed the new particle $\mathbf{x}_p(t+1)$ using (4), we will find the best bounding box (image region) in $\mathbf{S}$ which has the maximum overlap with $\mathbf{x}_p(t+1)$, and then approximate $\mathbf{x}_p(t+1)$ with this bounding box.

E. Combining Discriminative Parts

One remaining issue is how to combine the visual information from different parts of the object. The most straightforward approach is to concatenate features of all parts into one large vector and train a single classifier. However, this assumes that all parts are equally important for object classification and recognition. Certainly, this is not true. Motivated by this observation, we propose to use adaptive weighting to combine different parts P_i in (1). Here, we apply the max-margin template selection method of [43], [44]. Intuitively, the weight β_i represents the importance level of part P_i for the recognition

Algorithm 1: TPPL greedy search to find optimal part location for recognition task

Input: input image I; initial part location $InitP_i$; part classifier $w_{i,n}$
Output: final part location $OptiP_i$ with its final classification score $score(C_n)$
Symbols: K is maximum iteration, N is number of classes; M is number of parts; i is parts index; $CurrP_i$ is the optimal part location in current iteration; $w_{i,n}$ is the part classifier for part i;

for *class* $n \leftarrow 1$ ***to*** N **do**
  for *part* $i \leftarrow 1$ ***to*** M **do**
    Initialization: $CurrP_i \leftarrow InitP_i$
    for *iteration* $k \leftarrow 1$ ***to*** K **do**
      $T_j \leftarrow$ candidate regions that have ≥ 0.5 overlap with it $CurrP_i$
      Select the current part region: $CurrP_i \leftarrow \arg\max_{T_j} \Phi_i(I, T_j) \times w_{i,n}$
      Compute $y(C_n|P_i) \leftarrow \Phi_i(I, CurrP_i) \times w_{i,n}$
    Set the new part region: $OptiP_i \leftarrow CurrP_i$
  Evaluate $score(C_n) \leftarrow \sum_{i=1}^{M} \beta_i \times y(C_n|OptiP_i)$
return *Final score* $score(C_n)$ *and optimal part regions* $OptiP_i$

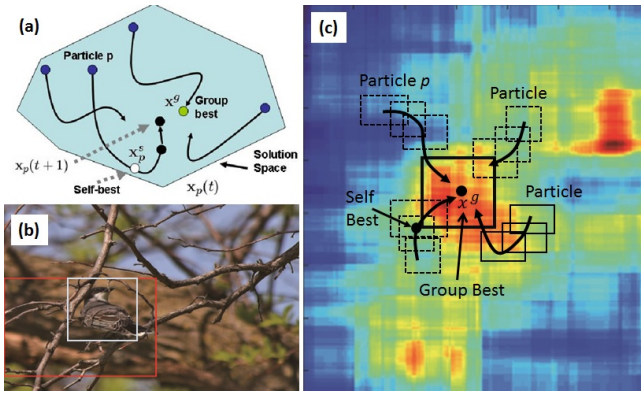


Fig. 4. Illustration of particle swarm optimization for progressive part localization.

task. So, for those parts which contain significant distinctive information, such as the head and body, they should have larger weights than others, such as legs, which are less useful. Let $\Phi_i(I, p_i)$ be the feature for part P_i in image I, and w_{i,C_n} be the weight for part P_i and class C_n. Our task is to learn the weights $\beta = \{\beta_1, \beta_2, ...\beta_M\}$ such that

$$\beta = arg\min_{\beta} \sum_{n=1}^{N} \sum_{c \neq C_n} \max(0, 1 - \beta^T u^n_{C_n,c})^2 + \lambda \|\beta\|_1, \quad (6)$$

where N is the number of categories, the i-th element of $u^n_{C_n,c}$ is the difference in decision values between incorrect class c and correct class C_n,

$$u^n_{C_n,c}(i) = (w_{i,C_n} - w_{i,c})^T \times \Phi_i(I, P_i). \quad (7)$$

This is equivalent to a one-class SVM (an SVM with only positive labels) with an L_2 loss and L_1 regularization, and can thus be solved effectively by standard SVM solvers. The final classification is given by

$$\arg\max_{n} score(C_n) = \sum_{i=1}^{M} \beta_i \Phi_i(I, P_i) \times w_{i,n}. \quad (8)$$

IV. Experimental Results

A. Experimental Setup

In this section, we evaluate and analyze the performance of our algorithm and compare it with state-of-the-art algorithms. We will use the Caltech-UCSD Birds-200-2011 (CUB-200-2011) dataset [42] which has been widely used in the literature for performance evaluations of fine-grained object recognition. The task is to classify 200 species of birds, which is challenging due to the strong similarity between different categories. For example, Fig. 5 shows some sample images which are even difficult for human to recognize them accurately. Fig. 5(a) shows samples of the same class with different occlusion, pose variation and clutter background. They have very large intra-class variations. In (b), each row contains samples from the same classes. They have very strong inter-class ambiguity.

In the dataset, each image is labeled with its species and with the bounding box for the whole bird. It also provides at most 15 landmark points for bird parts. We trained and tested on the sample split settings provided by the dataset, which contains around 30 training samples for each species. In our experiment, two types of semantic templates, i.e., "head" and "body" are defined, as in [24], [25]. Because there is no such annotation, we follow the same procedure in [24] to obtain the corresponding rectangles covering annotated parts distributed within bird heads and bodies. During our tests, no ground truth bounding box is required for part localization or key point prediction.

We use the Part-based SPP-net as our baseline method. We first present results to demonstrate the capability of our progressive Part-SPP in accurate part localization, then compare its fine-grained classification performance with state-of-the-art methods, demonstrating how our task-driven progressive part localization framework can significantly improve the classification accuracy. We used the open-source package Caffe [45] to extract deep features and fine-tune our CNNs.

B. Part Localization

For the part localization, we first analyze the detection error of individual body parts and compared our TPPL with other

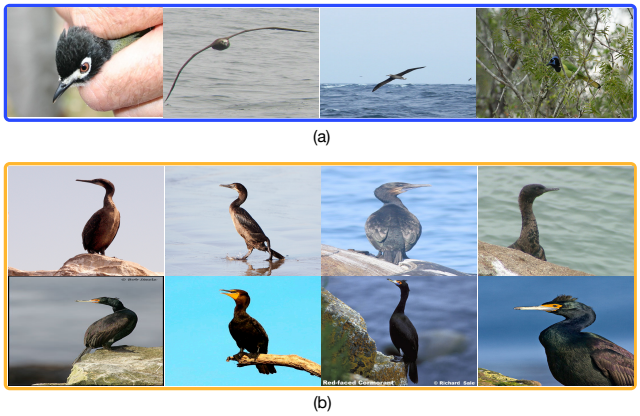


Fig. 5. CUB-200-2011 is a very challenging fine-grained bird dataset. (a) sample images from the same class with large intra-class variations and (b) samples from different classes in each row. It demonstrates the high-degree of similarity between subcategories.

TABLE II
COMPARISON WITH STATE-OF-THE-ART METHODS IN TERMS OF PART LOCALIZATION ACCURACY ON THE CUB-200-2011 DATASET. PART LOCALIZATION WAS PERFORMED WITHOUT THE GROUND-TRUTH BOUNDING BOXES.

Methods		head	body
Strong DPM [46]		37.44%	47.08%
P-RCNN [24]	Δ_{box}	60.56%	65.31%
	$\Delta_{geometric}$ with δ^{MG}	61.94%	70.16%
	$\Delta_{geometric}$ with δ^{NP}	61.42%	70.68%
Part-SPP	Δ_{box}	79.89%	80.31%
	$\Delta_{geometric}$ with δ^{MG}	81.62%	82.29%
	$\Delta_{geometric}$ with δ^{NP}	82.83%	83.43%
TPPL	Δ_{box}	80.15%	80.88%
	$\Delta_{geometric}$ with δ^{MG}	81.77%	82.69%
	$\Delta_{geometric}$ with δ^{NP}	**83.52%**	**84.01%**

state-of-the-art methods. Results in Table II are provided in terms of the Percentage of Correctly Localized Parts (PCP) metric. Here, Δ_{box} is the box constraint proposed in [24]. $\Delta_{geometric}$ with δ^{MG} means that parts are under the mixture of Gaussian geometric constraints, and $\Delta_{geometric}$ with δ^{NP} denotes the nearest neighbor geometric constraints.

In our experiment setting, no ground truth bounding boxes or part annotations are given during testing, and the task is to correctly localize the whole object bounding box. Detected body parts which have more than $\geq 50\%$ overlapping with the ground truth are considered as correct. This is very important in practice because obtaining the ground truth bounding box during testing is very labor-intensive. Table II shows that our system can produce accurate part localization, even without any bounding box infomation. For the head parts, our best result is 83.52% against the previous 37.44% in [46] and 61.42% in [24]. It outperformed the state-of-the-art method by more than 20%. For the bird body, our accuracy is as high as 84.01%. Fig. 6 shows six pairs of detected parts (bird body and head) obtained by the Part-RCNN method [24] (top) and our method (bottom), both with the *nearest neighbor geometric constraint*. For the entire bounding box our, our best detection AP is 94.96% under *nearest neighbor geometric constraint*. We can see that our task-driven progressive part localization can correct the part localization error and achieves significantly improved classification performance. We also show some failure cases in Fig. 7.

C. Fine-grained Image Categorization

In the following, we present results on the standard fine-grained categorization task using the widely used CUB-200-2011 benchmark dataset. In first set of results shown in Table III, the ground truth bounding box is given during the test, as in most state-of-art methods. This makes the classification task somehow easier. Lin *et al.* [47] introduced deep LAC, which combines detector, normalizer and classifier into a unified network, achieving 80.26% classification accuracy. Here, the oracle method uses the ground truth bounding box and part annotations for both training and testing. The second set of results is in a more challenging setting where the bird bounding box is *unknown* during testing. As shown in Table III, we can see that even our baseline Part-SPP works much better than the state-of-the-art methods. Here, -ft means extracting deep features from fine-tuned CNN models using each semantic part. TPPL represents our task-driven part localization. The oracle method uses the ground truth bounding box and part annotations for both training and test time.

We have achieved 80.55% classification accuracy without task-driven progressive part localization, which surpasses the Pose Normalization [25] by nearly 5%. This means our Part-SPP can generate very promising detection results, and no bounding box are needed during testing. With the progressive task-driven part localization (denoted as **TPPL**), our method achieves 81.68% classification accuracy, which outperforms the best result in the first setting where the ground-truth bounding boxes are provided. This means that our algorithm outperforms previous state-of-the-art methods even without using the ground truth bounding box. We noticed that two recent published papers [48] and [49] utilizing end-to-end deep learning framework surpassed our results and achieved 84.1% classification accuracy, however they do not undermine the contribution of our work. Both spatial transformer networks [49] and Bilinear CNNs [48] are introducing new network structures, while our paper is an orthogonal approach and is expected to be applicable to these new network structures.

In order to better demonstrate the discriminative region search capability of our task-driven part localization method, we conduct experiments on classification using one single part. As shown in Table IV, with progressive part detection and joint recognition, our approach significantly outperforms the "Oracle" method, which uses the ground truth bounding box in both training and testing. We trained a linear SVM using deep features on all the methods, so the performance gap could only come from the difference of part localization results. The -ft is the result of extracting deep features from fine-tuned CNN models on localized parts. Part Localization is performed without a bounding box. For part "head", our method does not boost its classification accuracy much, because it already contained rich discriminative patches such as color of eye and size of berk. However for "body" parts, the performance

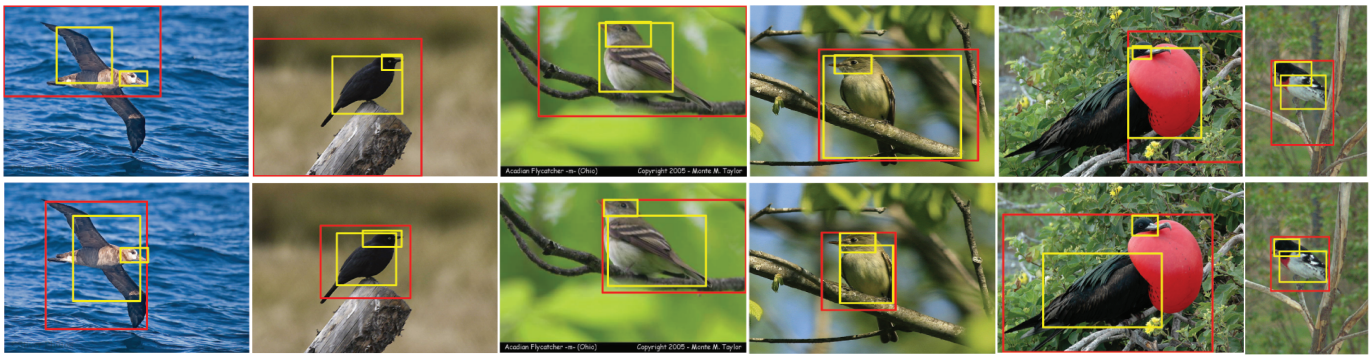


Fig. 6. Comparison of bird detection and part localization between Part-based RCNN [24] (top) and our Task-driven Part Localization(bottom), both using δ^{NP} constraint. This figure illuminates the excellent localization ability of our baseline method.

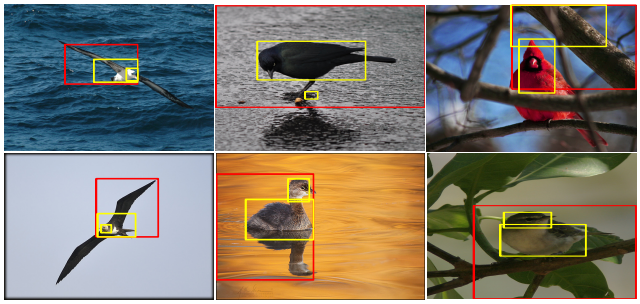


Fig. 7. Failure examples of our part localization (overlap with ground-truth < 0.5).

TABLE III
COMPARISON WITH THE STATE-OF-THE-ART METHODS ON THE CUB-200-2011 DATASET.

Fine-Grained Object Recognition Algorithms	Recognition Accuracy
Ground-Truth Bounding Box Provided	
POOF [1]	56.89%
Nonparametric Part Transfer [50]	57.84%
Symbiotic Segmentation [3]	59.40%
Alignment [22]	62.70%
DPD + DeCAF feature [23]	64.96%
Part-based RCNN [24]	76.37%
Deep LAC [47]	**80.26%**
Oracle	72.83%
Oracle-ft	82.14%
Ground-Truth Bounding Box **NOT** Provided	
DPD + DeCAF feature [23]	44.94%
Part-based RCNN [24]	73.89%
Pose Normalization [25]	75.70%
Part-SPP (This work)	80.55%
TPPL (This work)	**81.69%**

significantly improved by our method by almost 7%. This experiment clearly shows that our task-driven progressive part localization method can refine the detection results and automatically find the most discriminative patches for the recognition task.

D. Performance Gain Analysis

In this section, we study how different components of our algorithm contribute to the large performance gain. We use the CUB-200-2011 dataset to evaluate different configurations

TABLE IV
FINE-GRAINED CATEGORIZATION RESULTS ON CUB200-2011 BIRD DATASET WITH *only one part*.

Methods	Whole Object	Head	Body
Strong DPM [46]	38.02%		
R-CNN [36]	51.05%		
Part-RCNN [24]	62.75%		
Part-SPP (This Work)	72.23%	66.15%	63.70%
TPPL (This Work)	**73.90%**	**69.40%**	**70.87%**
Oracle-ft	73.01%	69.16%	64.36%

of components of our algorithm, including 1) different base detection algorithm usiing the SPP or RCNN; 2) different feature extraction networks with the ZF or AlexNet model; 3) different part combination strategy, discriminatively v.s. equally. We conducted a detailed and thorough analysis to answer that how much performance gain is obtained from every single design choice, and the analysis result is shown in Figure 8. By using SPP over RCNN as base detector, the classification accuracy is increased by 2.12%. Choosing the ZF network as feature extractor lead to 3.57% performance gain. Our discriminative parts combination boosted 0.97%, and finally the greedy search iteration can bring additional 1.14% performance gain. It should be noted that these different components are tightly coupled within the proposed algorithm, whose individual performance gain depends on the choices of other components.

E. Impact of CNN Architecture

According to recent paper [44], the choice of CNN architecture as feature extractor has a huge impact on the final categorization performance. In Krause et al.'s paper, by just replacing the AlexNet with the deeper and larger VGG-net, they get 8.3% increase in recognition accuracy, achieved state-of-the-art 82% on CUB-200-2011 dataset. Here we want to emphasis that our TPPL framework is also expected to be applicable to CNNs with deeper architectures like GoogleNet [38] and ResNet [51], and a performance boost can also be anticipated.

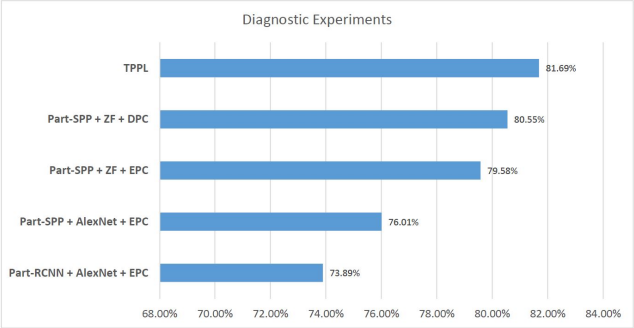


Fig. 8. Diagnostic Experiments on CUB-200-2011 Estimation to compare the performance of different part localization methods, feature extractor networks and parts combining methods. EPC is short for equally parts combination, DPC is short for discriminative parts combination

F. Discussion

Some interesting observations can be made from the above experiments. In the part localization stage, there is a significant performance gap between the Part-SPP methods with different constraints. For example, in Table II, for body detection, the performance gap is around 2%. But in the recognition stage, when classifying images just using features from body, the classification accuracy rate is almost the same, as shown in Table IV, the difference is only 0.02% between different methods. This suggests that, even if the initial detection results are poor, the task-driven part localization framework can find regions that contains more discriminative parts after a few iterations.

Second, we observe that the task-driven part localization method significantly improves the classification. When only a single part is detected and utilized for classification, as shown in Table IV, our TPPL boosts its performance significantly and outperforms those methods which use the ground-truth bounding boxes. This is mainly because the ground-truth bounding boxes are manually defined and may not be distinctive enough for classification. However, when all parts are used together for classification, as shown in Table III, the performance improvement is smaller. This is because the progressive localization procedure for different parts are different. Simply combining them together using weighted summation may not be the best solution and is worth further investigations.

V. Conclusions

In this paper, we have successfully developed a task-driven progressive part localization (TPPL) approach for fine-grained recognition. Our major finding is that the part detector should be jointly designed and progressively refined with the object classifier so that the detected parts can provide the most distinctive features for final object recognition. We start with a Part-based SPP-net (Part-SPP) as our baseline part detection, then develop a task-driven progressive part localization framework, which uses the trained object classifier to refine detection results. We then examine new image regions in the neighborhood using a particle swarm optimization approach, searching for more discriminative image regions which maximize the objective function with more distinctive clues for the recognition task. This procedure is performed in an iterative manner to progressively improve the joint part detection and object classification performance.

In future extension of this work, we will consider methods which can train the Part-SPP detectors in a weakly supervised setting, without any ground truth bounding boxes and part annotations. Given only the class label in training time, we hope to automatically discover and model parts, which are most distinctive for the accurate recognition.

References

[1] T. Berg and P. N. Belhumeur, "Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, 2013, pp. 955–962.

[2] S. Branson, C. Wah, F. Schroff, B. Babenko, P. Perona, and S. Belongie, "Visual recognition with humans in the loop," in *Computer Vision–ECCV 2010*, 2010, pp. 438–451.

[3] Y. Chai, V. Lempitsky, and A. Zisserman, "Symbiotic segmentation and part localization for fine-grained categorization," in *Computer Vision (ICCV), 2013 IEEE International Conference on*, 2013, pp. 321–328.

[4] R. Farrell, O. Oza, N. Zhang, V. Morariu, T. Darrell, and L. Davis, "Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, 2011, pp. 161–168.

[5] C. Wah, S. Branson, P. Perona, and S. Belongie, "Multiclass recognition and part localization with humans in the loop," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, 2011, pp. 2524–2531.

[6] N. Zhang, R. Farrell, and T. Darrell, "Pose pooling kernels for sub-category recognition," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012, pp. 3665–3672.

[7] J. Liu, A. Kanazawa, D. Jacobs, and P. Belhumeur, "Dog breed classification using part localization," in *Computer Vision–ECCV 2012*. Springer, pp. 172–185.

[8] O. Parkhi, A. Vedaldi, C. Jawahar, and A. Zisserman, "The truth about cats and dogs," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, 2011, pp. 1427–1434.

[9] O. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar, "Cats and dogs," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012, pp. 3498–3505.

[10] N. Kumar, P. N. Belhumeur, A. Biswas, D. W. Jacobs, W. J. Kress, I. C. Lopez, and J. V. Soares, "Leafsnap: A computer vision system for automatic plant species identification," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 502–516.

[11] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," *arXiv preprint arXiv:1306.5151*, 2013.

[12] L. Xie, J. Wang, B. Zhang, and Q. Tian, "Fine-grained image search," *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 636–647, 2015.

[13] O. Russakovsky, J. Deng, Z. Huang, A. Berg, and L. Fei-Fei, "Detecting avocados to zucchinis: What have we done, and where are we going?" in *Computer Vision (ICCV), 2013 IEEE International Conference on*, 2013, pp. 2064–2071.

[14] 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 *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, 2014.

[15] X. Zhou, K. Yu, T. Zhang, and T. S. Huang, "Image classification using super-vector coding of local image descriptors," in *Computer Vision–ECCV 2010*. Springer, 2010, pp. 141–154.

[16] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1704–1716, 2012.

[17] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *International Journal of computer vision*, vol. 105, no. 3, pp. 222–245, 2013.

[18] S. Bu, Z. Liu, J. Han, J. Wu, and R. Ji, "Learning high-level feature by deep belief networks for 3-D model retrieval and recognition," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2154–2167, 2014.

[19] Z. Guo and Z. Wang, "An unsupervised hierarchical feature learning framework for one-shot image recognition," *IEEE Transactions on Multimedia*, vol. 15, no. 3, pp. 621–632, 2013.

[20] N. Zhang, R. Farrell, F. Iandola, and T. Darrell, "Deformable part descriptors for fine-grained recognition and attribute prediction," in *Computer Vision (ICCV), 2013 IEEE International Conference on*, 2013, pp. 729–736.

[21] G. Chen, J. Yang, H. Jin, E. Shechtman, J. Brandt, and T. Han, "Selective pooling vector for fine-grained recognition," in *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, 2015, pp. 860–867.

[22] E. Gavves, B. Fernando, C. Snoek, A. Smeulders, and T. Tuytelaars, "Fine-grained categorization by alignments," in *Computer Vision (ICCV), 2013 IEEE International Conference on*, 2013, pp. 1713–1720.

[23] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2014.

[24] N. Zhang, J. Donahue, R. Girshick, and T. Darrell, "Part-based r-cnns for fine-grained category detection," in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 834–849.

[25] S. Branson, G. V. Horn, S. Belongie, and P. Perona, "Bird species categorization using pose normalized deep convolutional nets," in *British Machine Vision Conference (BMVC)*, Nottingham, 2014.

[26] A. Vedaldi, S. Mahendran, S. Tsogkas, S. Maji, R. Girshick, J. Kannala, E. Rahtu, I. Kokkinos, M. Blaschko, D. Weiss, B. Taskar, K. Simonyan, N. Saphra, and S. Mohamed, "Understanding objects in detail with fine-grained attributes," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2014, pp. 3622–3629.

[27] L. Bourdev, S. Maji, and J. Malik, "Describing people: A poselet-based approach to attribute classification," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, 2011, pp. 1543–1550.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[29] C. Huang and Z. He, "Task-driven progressive part localization for fine-grained recognition," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–9.

[30] B. Yao, G. Bradski, and L. Fei-Fei, "A codebook-free and annotation-free approach for fine-grained image categorization," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012, pp. 3466–3473.

[31] G. Martinez-Munoz, N. Larios, E. Mortensen, W. Zhang, A. Yamamuro, R. Paasch, N. Payet, D. Lytle, L. Shapiro, S. Todorovic, A. Moldenke, and T. Dietterich, "Dictionary-free categorization of very similar objects via stacked evidence trees," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 549–556.

[32] L. Xie, Q. Tian, R. Hong, S. Yan, and B. Zhang, "Hierarchical part matching for fine-grained visual categorization," in *Computer Vision (ICCV), 2013 IEEE International Conference on*, 2013, pp. 1641–1648.

[33] A. Angelova and P. Long, "Benchmarking large-scale fine-grained categorization," in *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on*, 2014, pp. 532–539.

[34] T. Berg, J. Liu, S. W. Lee, M. L. Alexander, D. W. Jacobs, and P. N. Belhumeur, "Birdsnap: Large-scale fine-grained visual categorization of birds," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2014, pp. 2019–2026.

[35] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[36] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2014, pp. 580–587.

[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[38] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*.

[39] Z. Wu, Y. Huang, and L. Wang, "Learning representative deep features for image set analysis," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1960–1968, 2015.

[40] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[41] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 818–833.

[42] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie, "The caltech-ucsd birds-200-2011 dataset," Tech. Rep., 2011.

[43] G. Chen, J. Yang, H. Jin, J. Brandt, E. Shechtman, A. Agarwala, and T. Han, "Large-scale visual font recognition," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2014, pp. 3598–3605.

[44] J. Krause, H. Jin, J. Yang, and L. Fei-Fei, "Fine-grained recognition without part annotations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5546–5555.

[45] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014, pp. 675–678.

[46] H. Azizpour and I. Laptev, "Object detection using strongly-supervised deformable part models," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 836–849.

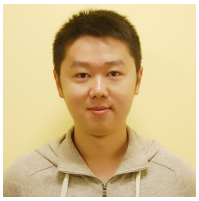
[47] D. Lin, X. Shen, C. Lu, and J. Jia, "Deep lac: Deep localization, alignment and classification for fine-grained recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1666–1674.

[48] T.-Y. Lin, A. RoyChowdhury, and S. Maji, "Bilinear cnn models for fine-grained visual recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1449–1457.

[49] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 2008–2016.

[50] C. Goering, E. Rodner, A. Freytag, and J. Denzler, "Nonparametric part transfer for fine-grained recognition," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 2014, pp. 2489–2496.

[51] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*, 2016.



Chen Huang received his B.Sc degree in electrical and computer engineering from Beihang University, China, in 2010. Now he is pursuing a Ph.D degree in Electrical and Computer Engineering from the University of Missouri-Columbia.

His research interests includes contour-based object detection, fine-grained image classification and convolutional neural network.

Zhihai He (S'98–M'01–SM'06) received the B.S. degree from Beijing Normal University, Beijing, China, and the M.S. degree from Institute of Computational Mathematics, Chinese Academy of Sciences, Beijing, China, in 1994 and 1997 respectively, both in mathematics, and the Ph.D. degree from University of California, Santa Barbara, CA, in 2001, in electrical engineering. In 2001, he joined Sarnoff Corporation, Princeton, NJ, as a Member of Technical Staff. In 2003, he joined the Department of Electrical and Computer Engineering, University of Missouri, Columbia. His current research interests include image/video processing and compression, network transmission, wireless communication, computer vision analysis, sensor networks, and embedded system design.

He received the 2002 IEEE Transactions on Circuits and Systems for Video Technology Best Paper Award and the SPIE VCIP Young Investigator Award in 2004. Currently, he has served as an Associate Editor for IEEE Transactions on Circuits and Systems for Video Technology, IEEE Transactions on Multimedia, and Journal of Visual Communication and Image Representation. He is also a guest-editor for IEEE TCSVT Special Issue on Video Surveillance. He is the Co-Chair of the 2007 International Symposium on Multimedia over Wireless in Hawaii. He is a member of the Visual Signal Processing and Communication Technical Committee of the IEEE Circuits and Systems Society, and serves as Technical Program Committee member or session chair of a number of international conferences.

Wenming Cao received his M.S. degree from The System Science Institute of China Science Academy, Beijing, China, and his Ph. D degree from School of Automation, Southeast University, Nanjing, China, in 1991 and 2003 respectively. During 2005 to 2007 he worked as post doc in the Semiconductors of China Science Academy.

Dr. Cao now serves as a professor in Shenzhen University with his research interests that include pattern recognition, image processing, and visual tracking. He has more than 80 publications on top-tier conferences and journals