Task-driven Progressive Part Localization for Fine-grained Recognition

Chen Huang Zhihai He

University of Missouri

chenhuang@mail.missouri.edu hezhi@missouri.edu

Abstract

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 classifies 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 start with a part-based SPP-net (Part-SPP) as our baseline part detector. We then develop 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, searching for more discriminative image regions to maximize 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.

1. Introduction

Fine-grained categorization, also referred to as subcategory recognition, is a rapidly growing subfield in object recognition. Different from traditional image classification such as scene or object recognition, fine-grained categorization deals with images which have subtle distinctions, such as recognizing the species of a bird [2, 6], the breed of a dog or cat [20], or the model of an aircraft [19]. Even in the ImageNet Challenge, an important issue in many state-of-the-art recognition algorithms is how to distinguish closely related subcategories [21]. Fine-grained categorization is a challenging computer vision problem since it needs to deal with subtle differences in the overall appearance between various classes (small inter-class variations) and large appearance variations in 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 [26].

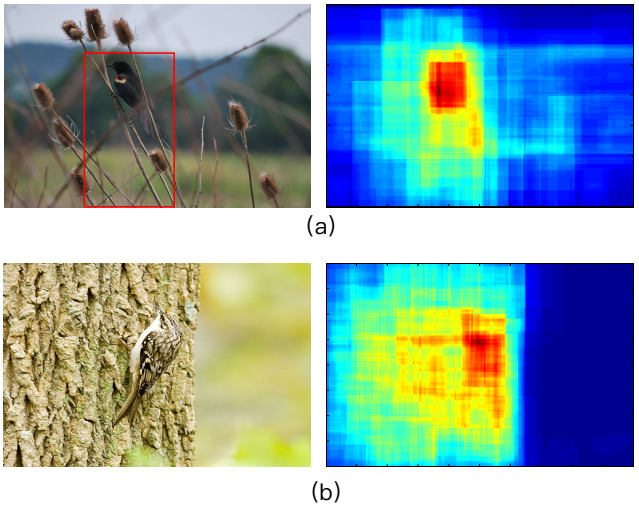

Figure 1: Samples to show the correlation between object part detection and classification. The left images are detection results. In these cases these detection are failed, as the target is either too small, or is too visually similar to the background. The images on the right are classification score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

Most visual object recognition methods are mainly based on global representation containing statistics of local features calculated in the whole image [14, 22]. However,

for fine-grained 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 [29, 2] 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 [28, 5, 8] have demonstrated significantly improved performance over earlier work [4].

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. Even if the two tasks are different, they are obviously 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 hardly visible objects. Fig. 1 shows two examples where the detection tasks failed; however, classification score response maps include a strong clue as to where the object is located. Moreover, we recognize that in some cases even the detected parts could be semantically accurate, it may not be optimal for classification, as these parts are subjectively defined by human input, and the part detection module is trained from these hand-labeled samples. We note that the feedback information from the classifier could be utilized to correct or refine these inaccurate or inefficient part detection results. So, in this work 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 subsequent 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 started with a revised version of Part-based R-CNN [28] and SPP-net [13] to generate initial detection results. We developed a greedy search 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 entire pipeline is shown in Fig. 2.

The major contributions of this paper are: (1) we recognize the correlation between accurate part localization and fine-grained categorization and the need for joint design of these two components. (2) We propose a task-driven progressive part localization(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 2, we review the related work on fine-grained object recognition. Section 3 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 4 and conclude with future work in Section 5.

2. Related work

Fine-grained image categorization has been extensively studied recently. Early work in this area concentrated on constructing discriminative global image representation [22]. It was later found suffering from the problem of losing subtle differences between subcategories. So, localizing semantic parts and extracting more discriminative features have become two mainstream aspects to boosting the recognition accuracy.

A. Semantic Part Localization. Incorporating precise part information has proved to be crucial in building accurate fine-grained recognition systems in recent studies. The Poselet [3] and DPM [9, 29] methods have previously been utilized to obtain part localizations with a modest degree of success. The state-of-the-art method of Part-based R-CNN [28] uses the R-CNN [11] approach for localizing parts, which 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 much faster SPP-net [13] as the baseline. With the part localization results of Part-based SPP-net as our initial guess, we refined it via an iterative search to find more discriminative regions for the later recognition task.

B. Feature Representation. The other aspect of related work is to represent regions of interest. Berg *et al.* has proposed the part-based one-vs-all features library POOF [2] as the mid-level features. Chai *et al.* [6] used Fisher vectors to learn global level and object part label representations. Deep convolutional neural network (CNN) has become the dominant approach for large-scale image clas-

sification since the work by Krizhevsky *et al.* [17] for ImageNet recognition with CNNs. The network structure of Krizhevsky *et al.* consists of five convolutional layers (*conv1* to *conv5*) with two fully-connected layers (*fc6* and *fc7*), 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 [23] 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.

3. Our Approach

In the following, we first introduce our baseline Part-SPP, a modified version of Part-based R-CNN by alternating the R-CNN with SPP-net for faster part detection and better performance. This is our baseline method. We then analyze the underlying assumption of a 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.

3.1. Adaptation of SPP-net for Part Detection

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 [11] 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 hard disk space to cache its intermediate features. To address this issue, we chosen the Spatial Pyramid Pooling (SPP-net) [13], which computes feature maps for entire images only once, and then pool features in arbitrary regions (sub-images) using a spatial pyramid pooling approach to generate fixed-length representations for part detection. He *et al.* [13] 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 first detects the semantic parts of objects, then extracts features from each localized regions, classifies each part independently, and finally combines all part classification scores to get the final result. The detector and classifier are designed independently. In our implementation, a SPP-net [13] is trained for the entire object and every part by treating each as a separate category, which means the same part of different subcategories 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 [24] 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 [13]. 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.* [28] 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 discriminatingly combined for final class label prediction, which will be discussed in Sec 3.4.

Instead of training a network from scratch, we use the ImageNet pre-trained Zeiler model in our experiments. This architecture is based on the Zeiler and Fergus (ZF) *fast* (smaller) model [27], 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 CUB-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 where $\mu = 0$ 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

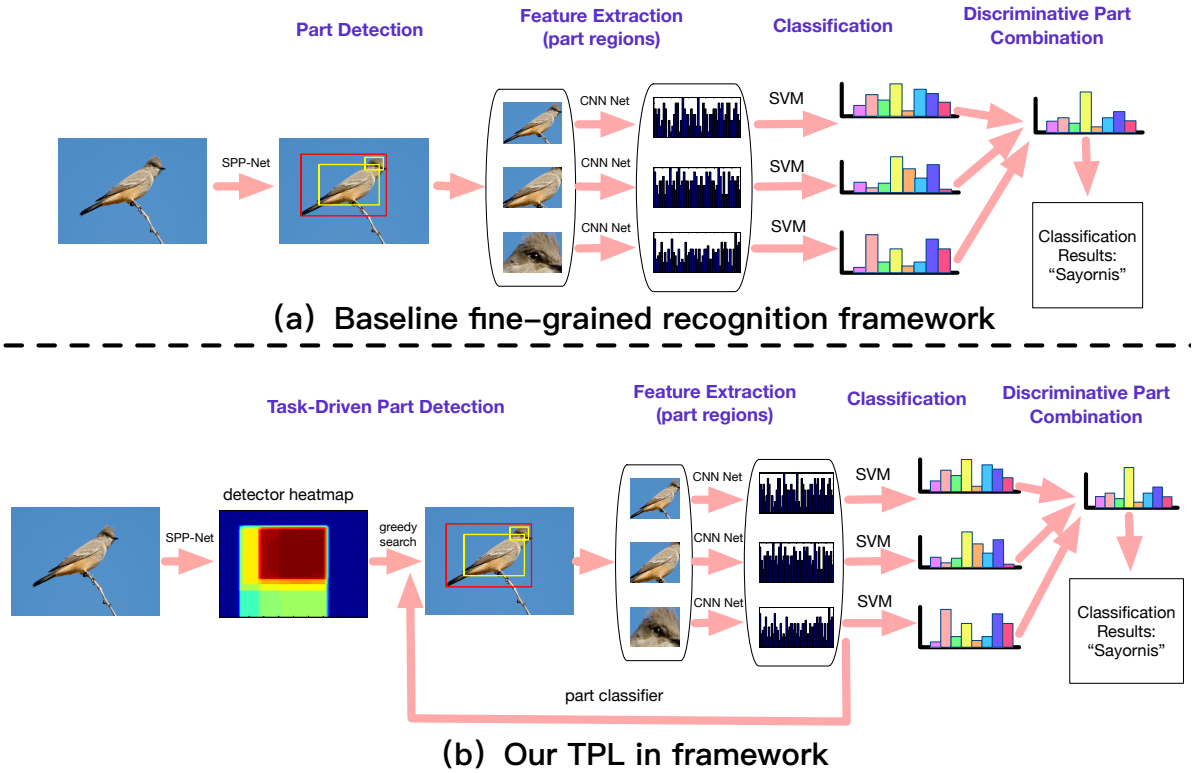

Figure 2: A conventional part-based recognition pipeline (upper figure) vs our proposed method pipeline (lower figure). In the task-driven part localization framework, part classifier is used to refine the part detector's results.

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.

3.2. 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 recognition performance. This leads to our task-driven progressive part localization(TPPL) method for object recognition.

Moreover, 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 narrative description, through our analysis, 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 classification performance?

We use the CUB-200-2011 dataset [25] to do a simple analysis. In this case, only object bounding boxes were detected as 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. In the whole dataset target objects in 5502 out of 5794 test im-

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

Table 1: Correlation between detection and classification from tests on the CUB-200-2011 dataset.

age were correctly detected in Table 1. Within these 5502 image, 3899 of them, about 70% were correctly classified while 30% were misclassified. This implies that the current part detection regions don't provide sufficiently accurate representation of the object for effective classification. In [28], 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 for the recognition task. 2) Currently, detection and classification are divided into two separate stages. 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.

3.3. Task-driven Progressive Part Localization

The overall greedy search process of our task-driven progressive part localization is shown in Fig. 3. As in [28], 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 (ROI) $\{R_1, R_2, ..., R_K\}$ are sampled on the test image using the so-called selective search method developed in [24]. 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 ≥ 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.

3.4. Combining Discriminative Parts

One remaining issue is how to combine the multiple parts of visual information. The most straightforward approach is to concatenate features at each part into one long vector and train a single classifier. However, doing so ignores the observation that not all parts are equally useful for recognition. Motivated by this, we are trying to obtain the weights β_i of each part for combining different parts P_i in (1). Here we apply the max-margin template selec-

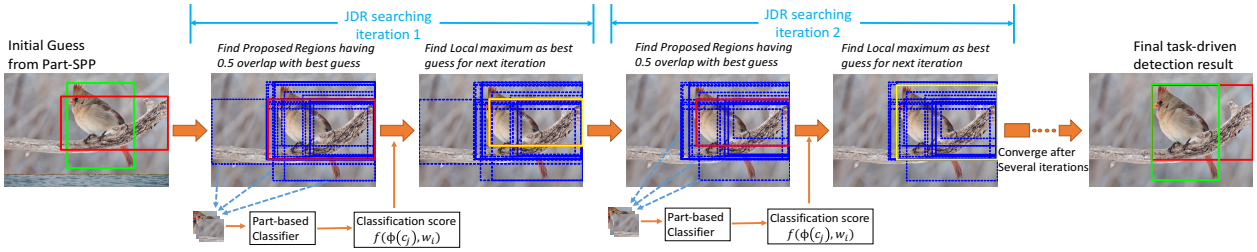

Figure 3: The overview of our task-driven part progressive localization(TPPL) framework. 1) The detection result of part-SPP is considered as an initial guess, shown as a red rectangle in the leftmost image. The green rectangle is the ground-truth bounding box. 2) We selected the proposed regions which 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.

tion method of [7, 16]. Intuitively, β_i represents the importance level of part P_i for the recognition task. So, for those parts which contain a lot of distinctive information, such as head and body, they should have larger weights than others, such as legs, which are rarely 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

$$\beta = arg\min_{\beta} \sum_{n=1}^{N} \sum_{c \neq C_n} \max(0, 1 - \beta^T u^n_{C_n,c})^2 + \lambda \|\beta\|_1, \tag{4}$$

where N is the number of categories, and 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,

$$u^n_{C_n,c}(i) = (w_{i,C_n} - w_{i,c})^T \times \Phi_i(I, P_i). \tag{5}$$

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

$$\arg\max_n score(C_n) = \sum_{i=1}^{M} \beta_i \Phi_i(I, P_i) \times w_{i,n}. \tag{6}$$

4. Experiment

4.1. Experimental setup

In this section, we conducted performance evaluations of our proposed method. Specifically, we tested on the widely-used Caltech-UCSD Birds-200-2011 (CUB-200-2011) dataset [25]. The task was to classify 200 species of birds and was challenging due to the high degree of similarity between different categories. For the sample images shown in Fig. 4, it is even difficult for a bird expert to recognize them correctly. Fig. 4(a) shows samples with different occlusion, pose variation and clutter background. In (b), each row contains samples from the same classes. This shows very large intra-class variations and strong inter-class ambiguity.

In the dataset, each image is labeled with its species and with the bounding box for the whole bird. We trained and tested on the sample splits settings provided by the dataset, which contains around 30 training samples for each species. In our experiment, two kinds of semantic parts, i.e., "head" and "body" are defined, as in [28, 5]. Because there is no such annotation, we follow the same procedure in [28] 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.

In this experiment, we first present results to demonstrate the ability of our progressive Part-SPP to accurately localize parts, 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 classification accuracy. We used the open-source package Caffe [15] to extract deep features and fine-tune our CNNs.

4.2. Part Localization

For part localization, we first analyzed the detection error of individual parts and compared our TPPL with other state-of-the-art methods. Results in Table 2 are provided in terms of the Percentage of Correctly Localized Parts (PCP) metric. Here Δ_{box} is the box constraint developed in [28]. $\Delta_{geometric}$ 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 were given during testing, and the task was to correctly localize whole object bounding box, with parts having $\geq 50\%$ overlap with ground truth considered as correct. This is very important in practice because the ground

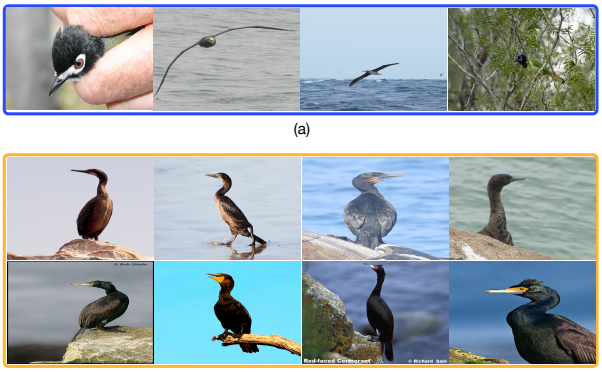

Figure 4: CUB-200-2011 is a very challenging fine-grained bird dataset. (a) sample images from the same class with large intra-class variations and (b) samples from different classes in each row. It demonstrates the high-degree of similarity between subcategories.

truth bounding box during testing is very labor-intensive to obtain. Table 2 show that our system can produce accurate part localization, even without any bounding box information. For the head parts, our best result was 83.52% against the previous 37.44% in [1] and 61.42% in [28], whicch outperformed the state-of-the-art methods by more than 20%. For the bird bodies, our accuracy was as high as 84.01%. Fig. 5 shows six pairs of detected parts (bird body and head) obtained by the Part-RCNN method [28] (top) and our method (bottom), both with *nearest neighbor geometric constraint*. We can see that our task-driven progressive part localization can correct the part localization error and achieve improved classification performance. We also show some failure cases in Fig. 6.

Methods		head	body
Strong DPM [1]		37.44%	47.08%
P-RCNN	Δ_{box}	60.56%	65.31%
	$\Delta_{geometric}$ with δ^{MG}	61.94%	70.16%
	$\Delta_{geometric}$ with δ^{NP}	61.42%	70.68%
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%**

Table 2: Comparison of 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.

4.3. 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. In first set of results shown in Table 3, 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.* [18] introduced deep LAC, which combines detector, normalizer and classifier into a unified network, achieving 80.26% classification accuracy. 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 3, 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 progressive part localization.

We achieved 80.55% classification accuracy without task-driven progressive part localization, which almost surpasses the Pose Normalization [5] by 5%. 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. So we can assert that our algorithm outperforms previous state-of-the-art methods even without using the ground truth bounding box.

Algorithms	Recognition Accuracy
Ground-Truth Bounding Box Provided	
POOF [2]	56.89%
Nonparam Part Transfer [12]	57.84%
Symbiotic Segmentation [6]	59.40%
Alignment [10]	62.70%
DPD + DeCAF feature [8]	64.96%
Part-based RCNN [28]	76.37%
Deep LAC [18]	**80.26%**
Oracle	72.83%
Oracle-ft	82.14%
Ground-Truth Bounding Box **NOT** Provided	
DPD + DeCAF feature [8]	44.94%
Part-based RCNN [28]	73.89%
Pose Normalization [5]	75.70%
Part-SPP-ft (This work)	80.55%
Part-SPP-ft + TPPL (This work)	**81.69%**

Table 3: Comparison of the state-of-the-art methods on the CUB-200-2011 dataset.

In order to better demonstrate the discriminative region search capability of our TPPL method, we conducted experiments on classifications using one single part. As show in Table 4, with TPPL, our approach even outperformed the "Oracle" method, which uses the ground truth bounding box in both training and testing. We trained a linear

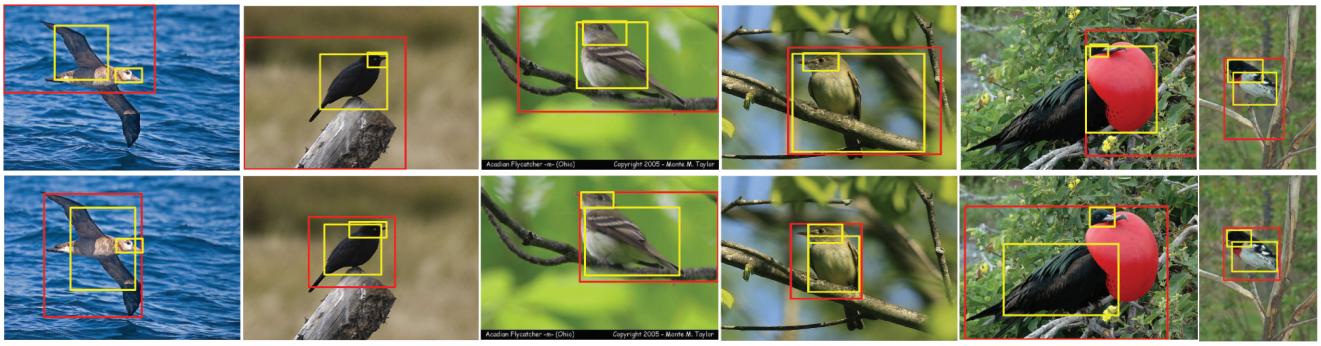

Figure 5: Comparison of bird detection and part localization between Part-based RCNN [28] (top) and our Task-driven Part Localization(bottom), both using δ^{NP} constraint. This figure illuminates the excellent localization ability of our baseline method.

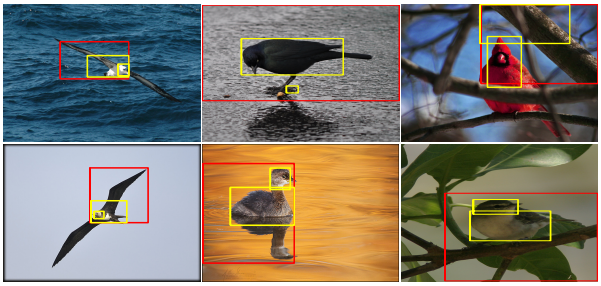

Figure 6: Failure examples of our part localization (overlap with ground-truth < 0.5).

Methods	Whole box	Head	Body
Strong DPM [1]	38.02%		
R-CNN [11]	51.05%		
Part-RCNN [28]	62.75%		
Part-SPP-ft	72.23%	66.15%	63.70%
Part-SPP-ft + TPPL	**73.90%**	**69.40%**	**70.87%**
Oracle-ft	73.01%	69.16%	64.36%

Table 4: Fine-grained categorization results on CUB200-2011 bird dataset with *only one part*.

SVM using deep features on all the methods, so the performance gap could only come from the difference of part localization results. The -ft desination is the result of extracting deep features from fine-tuned CNN models on localized parts. Part Localization was performed without a bounding box. For part "head", our method did not boost its classification accuracy much, because it already contained rich discriminative patches such as color of eye and size of beak. However for "body" parts, the performance was 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.

4.4. Discussion

Some interesting observations can be made from the above experiments. When only a single part is detected and utilized for classification, as shown in Table 4, our TPPL boosts its performance significantly and outperforms those methods which use ground-truth bounding boxes. This is mainly because 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 3, the performance improvement is smaller. This is because the progress 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.

5. Conclusion

This paper traces our development of 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 started with a Part-based SPP-net (Part-SPP) as our baseline part detection, then developed a task-driven progressive part localization framework, which took the predicted boxes of Part-SPP as an initial guess, then examines new image regions in the neighborhood, searching for more discriminative parts which maximized the recognition task. This procedure was performed in an iterative manner to progressively improve the joint part detection and object classification performance. In future extensions of this work, we will consider methods which can train the part detectors in a weakly supervised setting without any ground truth bounding box labeling or part annotations.

References

[1] H. Azizpour and I. Laptev. Object detection using strongly-supervised deformable part models. In *Computer Vision–ECCV 2012*, pages 836–849. Springer, 2012.

[2] 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*, pages 955–962, 2013.

[3] 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*, pages 1543–1550, 2011.

[4] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3D human pose annotations. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1365–1372, 2009.

[5] 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.

[6] 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*, pages 321–328, 2013.

[7] 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*, pages 3598–3605, 2014.

[8] 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.

[9] 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*, 32(9):1627–1645, 2010.

[10] 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*, pages 1713–1720, 2013.

[11] 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*, pages 580–587, 2014.

[12] 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*, pages 2489–2496, 2014.

[13] 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*, 37(9):1904–1916, 2015.

[14] 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*, 34(9):1704–1716, 2012.

[15] 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*, pages 675–678. ACM, 2014.

[16] 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*, pages 5546–5555, 2015.

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

[18] 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*, pages 1666–1674, 2015.

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

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

[21] 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*, pages 2064–2071, 2013.

[22] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *International Journal of computer vision*, 105(3):222–245, 2013.

[23] 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*.

[24] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[25] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical report, 2011.

[26] 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.

[27] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014*, pages 818–833. Springer, 2014.

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

[29] 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*, pages 729–736, 2013.