Part-Based R-CNNs
for Fine-Grained Category Detection

Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell

University of California, Berkeley, USA
{nzhang,jdonahue,rbg,trevor}@eecs.berkeley.edu

Abstract. Semantic part localization can facilitate fine-grained categorization by explicitly isolating subtle appearance differences associated with specific object parts. Methods for pose-normalized representations have been proposed, but generally presume bounding box annotations at test time due to the difficulty of object detection. We propose a model for fine-grained categorization that overcomes these limitations by leveraging deep convolutional features computed on bottom-up region proposals. Our method learns whole-object and part detectors, enforces learned geometric constraints between them, and predicts a fine-grained category from a pose-normalized representation. Experiments on the Caltech-UCSD bird dataset confirm that our method outperforms state-of-the-art fine-grained categorization methods in an end-to-end evaluation without requiring a bounding box at test time.

Keywords: Fine-grained recognition, object detection, convolutional models.

1   Introduction

The problem of visual fine-grained categorization can be extremely challenging due to the subtle differences in the appearance of certain parts across related categories. In contrast to basic-level recognition, fine-grained categorization aims to distinguish between different breeds or species or product models, and often requires distinctions that must be conditioned on the object pose for reliable identification. Facial recognition is the classic case of fine-grained recognition, and it is noteworthy that the best facial recognition methods jointly discover facial landmarks and extract features from those locations.

Localizing the parts in an object is therefore central to establishing correspondence between object instances and discounting object pose variations and camera view position. Previous work has investigated part-based approaches to this problem [7, 16, 29, 43, 46]. The bottleneck for many pose-normalized representations is indeed accurate part localization. The Poselet [8] and DPM [17] methods have previously been utilized to obtain part localizations with a modest degree of success; methods generally report adequate part localization only when given a known bounding box at test time [11, 20, 35, 36, 42]. By developing a novel deep part detection scheme, we propose an end-to-end fine grained categorization system which requires no knowledge of object bounding box at test time,
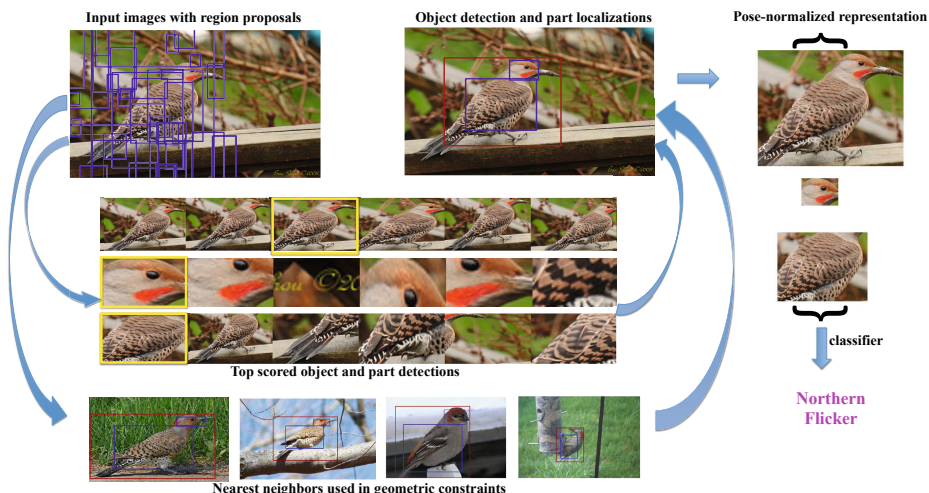
Fig. 1. Overview of our part localization Starting from bottom-up region proposals (top-left), we train both object and part detectors based on deep convolutional features. During test time, all the windows are scored by all detectors (middle), and we apply non-parametric geometric constraints (bottom) to rescore the windows and choose the best object and part detections (top-right). The final step is to extract features on the localized semantic parts for fine-grained recognition for a pose-normalized representation and then train a classifier for the final categorization. Best viewed in color.

and can achieve performance rivaling previously reported methods requiring the ground truth bounding box at test time to filter false positive detections.

The recent success of convolutional networks, like [26], on the ImageNet Challenge [22] has inspired further work on applying deep convolutional features to related image classification [14] and detection tasks [21]. In [21], Girshick et al. achieved breakthrough performance on object detection by applying the CNN of [26] to a set of bottom-up candidate region proposals [40], boosting PASCAL detection performance by over 30% compared to the previous best methods. Independently, OverFeat [37] proposed localization using a CNN to regress to object locations. However, the progress of leveraging deep convolutional features is not limited to basic-level object detection. In many applications such as fine-grained recognition, attribute recognition, pose estimation, and others, reasonable predictions demand accurate part localization.

Feature learning has been used for fine-grained recognition and attribute estimation, but was limited to engineered features for localization. DPD-DeCAF [47] used DeCAF [14] as a feature descriptor, but relied on HOG-based DPM [17] for part localization. PANDA [48] learned part-specific deep convolutional networks whose location was conditioned on HOG-based poselet models. These models lack the strength and detection robustness of R-CNN [21]. In this work we explore a unified method that uses the same deep convolutional representation for detection as well as part description.

We conjecture that progress made on bottom-up region proposal methods, like selective search [40], could benefit localization of smaller parts in addition to whole objects. As we show later, average recall of parts using selective search proposals is 95% on the Caltech-UCSD bird dataset.

In this paper, we propose a part localization model which overcomes the limitations of previous fine-grained recognition systems by leveraging deep convolutional features computed on bottom-up region proposals. Our method learns part appearance models and enforces geometric constraints between parts. An overview of our method is shown in Figure 1. We have investigated different geometric constraints, including a non-parametric model of joint part locations conditioned on nearest neighbors in semantic appearance space. We present state-of-the-art results evaluating our approach on the widely used fine-grained benchmark Caltech-UCSD bird dataset [41].

2   Related Work

2.1   Part-Based Models for Detection and Pose Localization

Previous work has proposed explicit modeling of object part appearances and locations for more accurate recognition and localization. Starting with pictorial structures [18, 19], and continuing through poselets [8] and related work, many methods have jointly localized a set of geometrically related parts. The deformable parts model (DPM) [17], until recently the state-of-the-art PASCAL object detection method, models parts with additional learned filters in positions anchored with respect to the whole object bounding box, allowing parts to be displaced from this anchor with learned deformation costs. The "strong" DPM [3] adapted this method for the strongly supervised setting in which part locations are annotated at training time. A limitation of these methods is their use of weak features (usually HOG [12]).

2.2   Fine-Grained Categorization

Recently, a large body of computer vision research has focused on the fine-grained classification problem in a number of domains, such as animal breeds or species [16, 25, 30, 32, 36, 45], plant species [1, 2, 5, 33, 34, 38], and man-made objects [31, 39].

Several approaches are based on detecting and extracting features from certain parts of objects. Farrell et al. [16] proposed a pose-normalized representation using poselets [8]. Deformable part models [17] were used in [36, 47] for part localization. Based on the work of localizing fiducial landmarks on faces [6], Liu et al. [30] proposed an exemplar-based geometric method to detect dog faces and extract highly localized image features from keypoints to differentiate dog breeds. Furthermore, Berg et al. [7] learned a set of highly discriminative intermediate features by learning a descriptor for each pair of keypoints. Moreover, in [29], the authors extend the non-parametric exemplar-based method of [6] by enforcing pose and subcategory consistency. Yao et al. [44] and Yang et al. [43] have

investigated template matching methods to reduce the cost of sliding window approaches. All these part-based methods, however, require the ground truth bounding box at test time for part localization or keypoint prediction.

Human-in-the-loop methods [9, 13, 15] ask a human to name attributes of the object, click on certain parts or mark the most discriminative regions to improve classification accuracy. Segmentation-based approaches are also very effective for fine-grained recognition. Approaches such as [11, 20, 35, 36, 42] used region-level cues to infer the foreground segmentation mask and to discard the noisy visual information in the background. Chai et al. [10] showed that jointly learning part localization and foreground segmentation together can be beneficial for fine-grained categorization. Similar to most previous part-based approaches, these efforts require the ground truth bounding box to initialize the segmentation seed. In contrast, the aim of our work is to perform end-to-end fine-grained categorization with no knowledge at test time of the ground truth bounding box. Our part detectors use convolutional features on bottom-up region proposals, together with learned non-parametric geometric constraints to more accurately localize object parts, thus enabling strong fine-grained categorization.

2.3   Convolutional Networks

In recent years, convolutional neural networks (CNNs) have been incorporated into a number of visual recognition systems in a wide variety of domains. At least some of the strength of these models lies in their ability to *learn* discriminative features from raw data inputs (e.g., image pixels), in contrast to more traditional object recognition pipelines which compute hand-engineered features on images as an initial preprocessing step. CNNs were popularized by LeCun and colleagues who initially applied such models to digit recognition [27] and OCR [28] and later to generic object recognition tasks [23]. With the introduction of large labeled image databases [22] and GPU implementations used to efficiently perform the massive parallel computations required for learning and inference in large CNNs, these networks have become the most accurate method for generic object classification [26].

Most recently, generic object detection methods have begun to leverage deep CNNs and outperformed any competing approaches based on traditional features. OverFeat [37] uses a CNN to regress to object locations in a coarse sliding-window detection framework. Of particular inspiration to our work is the R-CNN method [21] which leverages features from a deep CNN in a region proposal framework to achieve unprecedented object detection results on the PASCAL VOC dataset. Our method generalizes R-CNN by applying it to model object parts in addition to whole objects, which our empirical results will demonstrate is essential for accurate fine-grained recognition.

3   Part-Based R-CNNs

While [21] demonstrated the effectiveness of the R-CNN method on a generic object detection task (PASCAL VOC), it did not explore the application of this

method to simultaneous localization and fine-grained recognition. Because our work operates in this regime, we extend R-CNN to detect objects and localize their parts under a geometric prior. With hypotheses for the locations of individual semantic parts of the object of interest (e.g., the location of the head for an animal class), it becomes reasonable to model subtle appearance differences which tend to appear in locations that are roughly fixed with respect to these parts.

In the R-CNN method, for a particular object category, a candidate detection x with CNN feature descriptor $\phi(x)$ is assigned a score of $w_0^\intercal \phi(x)$, where w_0 is the learned vector of SVM weights for the object category. In our method, we assume a strongly supervised setting (e.g., [3]) in which at training time we have ground truth bounding box annotations not only for full objects, but for a fixed set of semantic parts $\{p_1, p_2, ..., p_n\}$ as well.

Given these part annotations, at training time all objects and each of their parts are initially treated as independent object categories: we train a one-versus-all linear SVM on feature descriptors extracted over region proposals, where regions with ≥ 0.7 overlap with a ground truth object or part bounding box are labeled as positives for that object or part, and regions with ≤ 0.3 overlap with any ground truth region are labeled as negatives. Hence for a single object category we learn whole-object ("root") SVM weights w_0 and part SVM weights $\{w_1, w_2, ..., w_n\}$ for parts $\{p_1, p_2, ..., p_n\}$ respectively. At test time, for each region proposal window we compute scores from all root and part SVMs. Of course, these scores do not incorporate any knowledge of how objects and their parts are constrained geometrically; for example, without any additional constraints the *bird head* detector may fire outside of a region where the *bird* detector fires. Hence our final joint object and part hypotheses are computed using the geometric scoring function detailed in the following section, which enforces the intuitively desirable property that pose predictions are consistent with the statistics of poses observed at training time.

3.1   Geometric Constraints

Let $X = \{x_0, x_1, \ldots, x_n\}$ denote the locations (bounding boxes) of object p_0 and n parts $\{p_i\}_{i=1}^n$, which are annotated in the training data, but unknown at test time. Our goal is to infer both the object location and part locations in a previously unseen test image. Given the R-CNN weights $\{w_0, w_1, \ldots, w_n\}$ for object and parts, we will have the corresponding detectors $\{d_0, d_1, \ldots, d_n\}$ where each detector score is $d_i(x) = \sigma(w_i^\intercal \phi(x))$, where $\sigma(\cdot)$ is the sigmoid function and $\phi(x)$ is the CNN feature descriptor extracted at location x. We infer the joint configuration of the object and parts by solving the following optimization problem:

$$X^* = \arg\max_X \Delta(X) \prod_{i=0}^n d_i(x_i) \tag{1}$$

where $\Delta(X)$ defines a scoring function over the joint configuration of the object and root bounding box. We consider and report quantitative results on several configuration scoring functions Δ, detailed in the following paragraphs.

Box constraints. One intuitive idea to localize both the object and parts is to consider each possible object window and all the windows inside the object and pick the windows with the highest part scores. In this case, we define the scoring function

$$\Delta_{\text{box}}(X) = \prod_{i=1}^{n} c_{x_0}(x_i) \tag{2}$$

where

$$c_x(y) = \begin{cases} 1 \text{ if region } y \text{ falls outside region } x \text{ by at most } \epsilon \text{ pixels} \\ 0 \text{ otherwise} \end{cases} \tag{3}$$

In our experiments, we let $\epsilon = 10$.

Geometric constraints. Because the individual part detectors are less than perfect, the window with highest individual part detector scores is not always correct, especially when there are occlusions. We therefore consider several scoring functions to enforce constraints over the layout of the parts relative to the object location to filter out incorrect detections. We define

$$\Delta_{\text{geometric}}(X) = \Delta_{\text{box}}(X) \left(\prod_{i=1}^{n} \delta_i(x_i) \right)^{\alpha} \tag{4}$$

where δ_i is a scoring function for the position of the part p_i given the training data. Following previous work on part localization from, e.g. [4, 17, 19], we experiment with three definitions of δ:

- $\delta_i^{MG}(x_i)$ fits a mixture of Gaussians model with N_g components to the training data for part p_i. In our experiments, we set $N_g = 4$.
- $\delta_i^{NP}(x_i)$ finds the K nearest neighbors in appearance space to $\tilde{x}_0$, where $\tilde{x}_0 = \arg\max d_0(x_0)$ is the top-scoring window from the root detector. We then fit a Gaussian model to these K neighbors. In our experiments, we set $K = 20$. Figure 2 illustrates some examples of nearest neighbors.

The DPM [17] models deformation costs with a per-component Gaussian prior. R-CNN [21] is a single-component model, motivating the δ^{MG} or δ^{NP} definitions. Our δ^{NP} definition is inspired by Belhumeur et al. [4], but differs in that we index nearest neighbors on appearance rather than geometry.

3.2   Fine-Grained Categorization

We extract semantic features from localized parts as well as the whole object. The final feature representation is $[\phi(x_0) \dots \phi(x_n)]$ where x_0 and $x_{1\dots n}$ are whole-object and part location predictions inferred using one of the models from the previous section and $\phi(x_i)$ is the feature representation of part x_i.
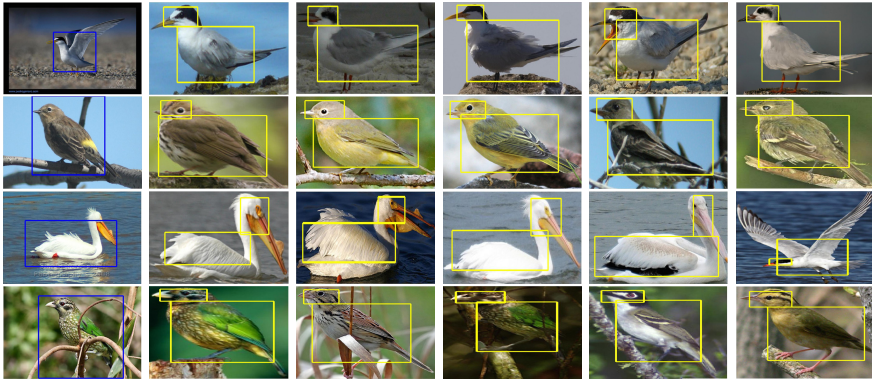
Fig. 2. Illustration of geometric constant δ^{NP}. In each row, the first column is the test image with an R-CNN bounding box detection, and the rest are the top-five nearest neighbors in the training set, indexed using `pool5` features and cosine distance metric.

In one set of experiments, we extract deep convolutional features $\phi(x_i)$ from an ImageNet pre-trained CNN, similar to DeCAF [14]. In order to make the deep CNN-derived features more discriminative for the target task of fine-grained bird classification, we also fine-tune the ImageNet pre-trained CNN for the 200-way bird classification task from ground truth bounding box crops of the original CUB images. 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 with $\mu = 0$ and $\sigma = 0.01$. We set fine-tuning learning rates as proposed by R-CNN [21], 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 of 10 times the global learning rate. For the whole object bounding box and each of the part bounding boxes, we independently finetune the ImageNet pre-trained CNN for classification on ground truth crops of each region warped to the 227×227 network input size, always with 16 pixels on each edge of the input serving as context as in R-CNN [21]. At test time, we extract features for the predicted whole object or part region using the network fine-tuned for that particular whole object or part.

For training the classifier, we employ a one-versus-all linear SVM using the final feature representation. For a new test image, we apply the whole and part detectors with the geometric scoring function to get detected part locations and use the features for prediction. If a particular part i was not detected anywhere in the test image (due to all proposals falling below the part detector's threshold, set to achieve high recall), we set its features $\phi(x_i) = \mathbf{0}$ (zero vector).

4   Evaluation

In this section, we present a comparative performance evaluation of our proposed method. Specifically, we conduct experiments on the widely-used fine-grained

benchmark Caltech-UCSD birds dataset [41] (CUB200-2011). The classification task is to discriminate among 200 species of birds, and is challenging for computer vision systems due to the high degree of similarity between categories. It contains 11,788 images of 200 bird species. Each image is annotated with its bounding box and the image coordinates of fifteen keypoints: the beak, back, breast, belly, forehead, crown, left eye, left leg, left wing, right eye, right leg, right wing, tail, nape and throat. We train and test on the splits included with the dataset, which contain around 30 training samples for each species. Following the protocol of [47], we use two semantic parts for the bird dataset: head and body.

We use the open-source package Caffe [24] to extract deep features and fine-tune our CNNs. For object and part detections, we use the Caffe reference model, which is almost identical to the model used by Krizhevsky et al. in [26]. We refer deep features from each layer as `conv`n, `pool`n, or `fc`n for the nth layer of the CNN, which is the output of a convolutional, pooling, or fully connected layer respectively. We use `fc6` to train R-CNN object and part detectors as well as image representation for classification. For δ^{NP}, nearest neighbors are computed using `pool5` and cosine distance metric.

4.1   Fine-Grained Categorization

We first present results on the standard fine-grained categorization task associated with the Caltech-UCSD birds dataset. The first set of results in Table 1 are achieved in the setting where the ground truth bounding box for the entire bird is known at test time, as most state-of-art methods assume, making the categorization task somewhat easier. In this setting, our part-based method with the local non-parametric geometric constraint δ^{NP} works the best without fine-tuning, achieving 68.1% classification accuracy without fine-tuning. Fine-tuning improves this result by a large margin, to over 76%. We compare our results against three state-of-the-art baseline approaches with results assuming the ground truth bounding box at test time. We use deep convolutional features as the authors of [14], but they use a HOG-based DPM as their part localization method. The increase in performance is likely due to better part localization (see Table 4). Oracle method uses the ground truth bounding box and part annotations for both training and test time.

The second set of results is in the less artificial setting where the bird bounding box is *unknown* at test time. Most of the literature on this dataset doesn't report performance in this more difficult, but more realistic setting. As Table 1 shows, in this setting our part-based method works much better than the baseline DPD model. We achieve 66.0% classification accuracy without finetuning , almost as good as the accuracy we can achieve when the ground truth bounding box is given. This means there is no need to annotate any box during test time to classify the bird species. With finetuned CNN models, our method achieves 73.89% classification accuracy. We are unaware of any other published results in this more difficult setting, but we note that our method outperforms previous state-of-the-art even without knowledge of the ground truth bounding box.

Table 1. Fine-grained categorization results on CUB200-2011 bird dataset. -ft means extracting deep features from finetuned CNN models using each semantic part. Oracle method uses the ground truth bounding box and part annotations for both training and test time.

Bounding Box Given	
DPD [47]	50.98%
DPD+DeCAF feature [14]	64.96%
POOF [7]	56.78%
Symbiotic Segmentation [10]	59.40%
Alignment [20]	62.70%
Oracle	72.83%
Oracle-ft	82.02%
Ours (Δ_{box})	67.55%
Ours ($\Delta_{\mathrm{geometric}}$ with δ^{MG})	67.98%
Ours ($\Delta_{\mathrm{geometric}}$ with δ^{NP})	68.07%
Ours-ft (Δ_{box})	75.34%
Ours-ft ($\Delta_{\mathrm{geometric}}$ with δ^{MG})	**76.37%**
Ours-ft ($\Delta_{\mathrm{geometric}}$ with δ^{NP})	76.34%
Bounding Box Unknown	
DPD+DeCAF [14] with no bounding box	44.94%
Ours (Δ_{null})	64.57%
Ours (Δ_{box})	65.22%
Ours ($\Delta_{\mathrm{geometric}}$ with δ^{MG})	65.98%
Ours ($\Delta_{\mathrm{geometric}}$ with δ^{NP})	65.96%
Ours-ft (Δ_{box})	72.73%
Ours-ft ($\Delta_{\mathrm{geometric}}$ with δ^{MG})	72.95%
Ours-ft ($\Delta_{\mathrm{geometric}}$ with δ^{NP})	**73.89%**

Another interesting experiment we did is to remove the part descriptors by only looking at the image descriptors inside the predicted bounding box. By having geometric constraints over part locations relative to object location, our method is able to help localize the object. As Table 2 shows, our method outperforms a single object detector using R-CNN, which means the geometric constraints helps our method better localize the object window. The detection of strong DPM is not as accurate as our method, which explains the performance drop. The "oracle" method uses the ground truth bounding box and achieves 57.94% accuracy, which is still much lower than the method in Table 1 of using both image descriptors inside object and parts.

4.2 Part Localization

We now present results evaluating in isolation the ability of our system to accurately localize parts. Our results in Table 4 are given in terms of the Percentage of Correctly Localized Parts (PCP) metric. For the first set of results, the whole object bounding box is given and the task is simply to correctly localize the parts inside of this bounding box, with parts having ≥ 0.5 overlap with ground truth counted as correct.

Table 2. Fine-grained categorization results on CUB200-2011 bird dataset with *no parts*. We trained a linear SVM using deep features on all the methods. Therefore only the bounding box prediction is the factor of difference. -ft is the result of extracting deep features from fine-tuned CNN model on bounding box patches.

Oracle (ground truth bounding box)	57.94%
Oracle-ft	68.29%
Strong DPM [3]	38.02%
R-CNN [21]	51.05%
Ours (Δ_{box})	50.17%
Ours ($\Delta_{\text{geometric}}$ with δ^{MG})	51.83%
Ours ($\Delta_{\text{geometric}}$ with δ^{NP})	52.38%
Ours-ft (Δ_{box})	62.13%
Ours-ft ($\Delta_{\text{geometric}}$ with δ^{MG})	62.06%
Ours-ft ($\Delta_{\text{geometric}}$ with δ^{NP})	**62.75%**

Table 3. Recall of region proposals produced by selective search methods on CUB200-2011 bird dataset. We use ground truth part annotations to compute the recall, as defined by the proportion of ground truth boxes for which there exists a region proposal with overlap at least 0.5, 0.6 and 0.7 respectively.

| Overlap	0.50	0.60	0.70
Bounding box	96.70%	97.68%	89.50%
Head	93.34%	73.87%	37.57%
Body	96.70%	85.97%	54.68%

For the second set of results, the PCP metric is computed on top-ranked parts predictions using the objective function described in Sec. 3.2. Note that in this more realistic setting we do not assume knowledge of the ground truth bounding box at test time – despite this limitation, our system produces accurate part localizations.

As shown in Table 4, for both settings of given bounding box and unknown bounding box, our methods outperform the strong DPM [3] method. Adding a geometric constraint δ^{NP} improves our results (79.82% for body localization compared to 65.42%). In the fully automatic setting, the top ranked detection and part localization performance on head is 65% better than the baseline method. $\Delta_{\text{null}} = 1$ is the appearance-only case with no geometric constraints applied. Although the fine-grained classification results don't show a big gap between $\Delta_{\text{geometric}}$ and Δ_{box}, we can see the performance gap for part localization. The reason for the small performance gap might be that deep convolutional features are invariant to small translations and rotations, limiting the impact of small localization errors on our end-to-end accuracy.

We also evaluate the recall performance of selective search region proposals [40] for bounding box and semantic parts. The results of recall given different overlapping thresholds are shown in Table 3. Recall for the bird head and body parts is high when the overlap requirement is 0.5, which provides the foundation for localizing these parts given the region proposals. However, we also observe

Table 4. Part localization accuracy in terms of PCP (Percentage of Correctly Localized Parts) on the CUB200-2011 bird dataset. There are two different settings: with given bounding box and without bounding box.

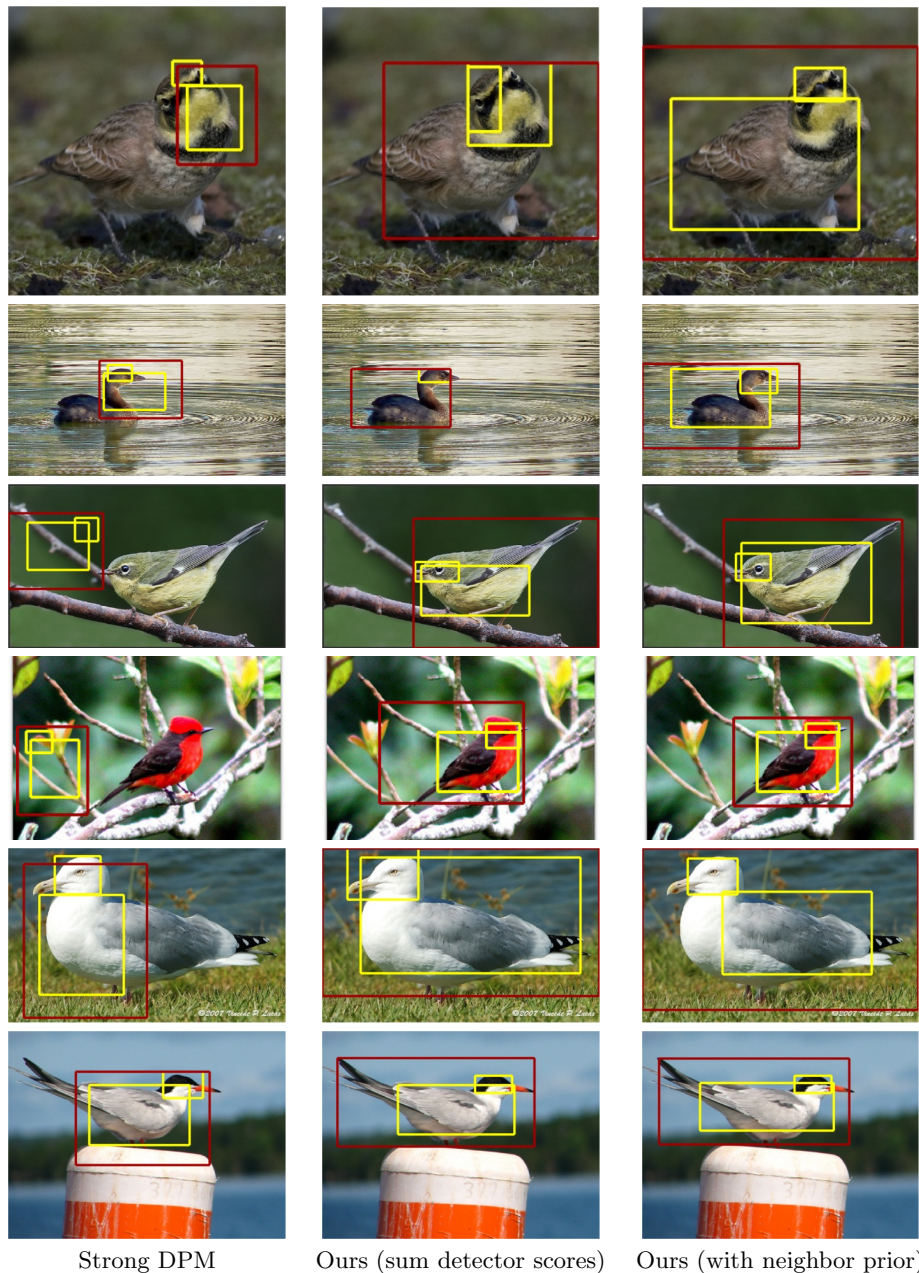| Bounding Box Given		
	Head	Body
Strong DPM [3]	43.49%	75.15%
Ours (Δ_{box})	61.40%	65.42%
Ours ($\Delta_{\text{geometric}}$ with δ^{MG})	66.03%	76.62%
Ours ($\Delta_{\text{geometric}}$ with δ^{NP})	**68.19%**	**79.82%**
Bounding Box Unknown		
	Head	Body
Strong DPM [3]	37.44%	47.08%
Ours (Δ_{null})	60.50%	64.43%
Ours (Δ_{box})	60.56%	65.31%
Ours ($\Delta_{\text{geometric}}$ with δ^{MG})	**61.94%**	70.16%
Ours ($\Delta_{\text{geometric}}$ with δ^{NP})	61.42%	**70.68%**
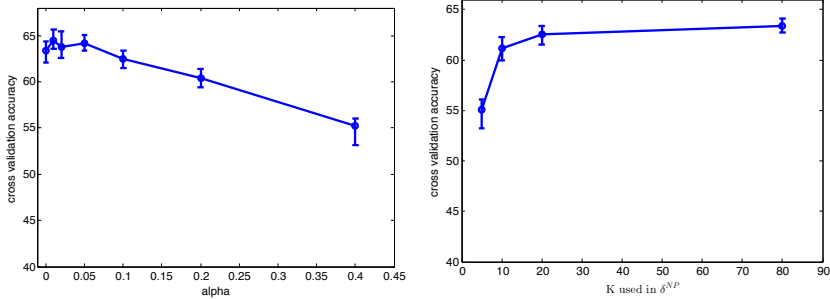
Fig. 3. Cross-validation results on fine-grained accuracy for different values of α (left) and K (right). We split the training data into 5 folds and use cross-validate each hyperparameter setting.

that the recall for head is below 40% when the overlap threshold is 0.7, indicating the bottom-up region proposals could be a bottleneck for precise part localization.

Other visualizations are shown in Figure 4. We show three detection and part localization for each image, the first column is the output from strong DPM, the second column is our methods with individual part predictions and the last column is our method with local prior. We used the model pretrained from [3] to get the results. We also show some failure cases of our method in Figure 5.

4.3   Component Analysis

To examine the effect of different values of α and K used in $\Delta_{\text{geometric}}$, we conduct cross-validation experiments. Results are shown in Figure 3. We fix $K = 20$ in

| Strong DPM | Ours (sum detector scores) | Ours (with neighbor prior) |

Fig. 4. Examples of bird detection and part localization from strong DPM [3] (left); our method using Δ_{box} part predictions (middle); and our method using δ^{NP} (right). All detection and localization results without any assumption of bounding box.
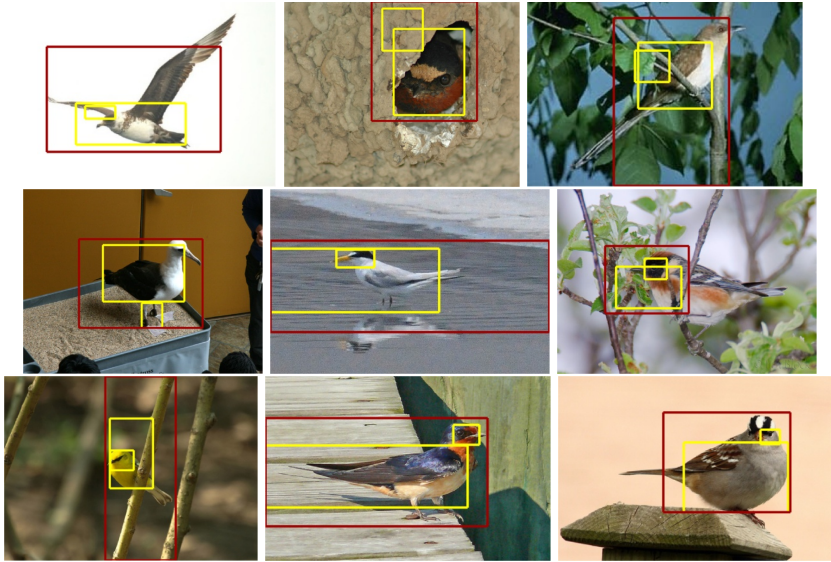
Fig. 5. Failure cases of our part localization using δ^{NP}

Figure 3, left and fix $\alpha = 0.1$ in Figure 3, right. All the experiments on conducted on training data in a cross-validation fashion and we split the training data into 5 folds. As the results show, the end-to-end fine-grained classification results are sensitive to the choice of α and $\alpha = 0$ is the case of Δ_{box} predictions without any geometric constraints. The reason why we have to pick a small α is the pdf of the Gaussian is large compared to the logistic score function output from our part detectors. On the other hand, the choice of K cannot be too small and it is not very sensitive when K is larger than 10.

5   Conclusion

We have proposed a system for joint object detection and part localization capable of state-of-the-art fine-grained object recognition. Our method learns detectors and part models and enforces learned geometric constraints between parts and with the object frame. Our experimental results demonstrate that even with a very strong feature representation and object detection system, it is highly beneficial to additionally model an object's pose by means of parts for the difficult task of fine-grained discrimination between categories with high semantic similarity. In future extensions of this work, we will consider methods which jointly model at training time the object category and each of its parts and deformation costs. We also plan to explore the weakly supervised setting in which we automatically discover and model parts as latent variables from only the object bounding box annotations. Finally, we will consider relaxing the use of selective search for smaller parts and employing dense window sampling.

References

1. Angelova, A., Zhu, S.: Efficient object detection and segmentation for fine-grained recognition. In: CVPR (2013)
2. Angelova, A., Zhu, S., Lin, Y.: Image segmentation for large-scale subcategory flower recognition. In: WACV (2013)
3. Azizpour, H., Laptev, I.: Object detection using strongly-supervised deformable part models. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part I. LNCS, vol. 7572, pp. 836–849. Springer, Heidelberg (2012)
4. Belhumeur, P.N., Jacobs, D., Kriegman, D., Kumar, N.: Localizing parts of faces using a consensus of exemplars. In: CVPR (2011)
5. Belhumeur, P.N., et al.: Searching the world's herbaria: A system for visual identification of plant species. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 116–129. Springer, Heidelberg (2008)
6. Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing parts of faces using a consensus of exemplars. In: CVPR (2011)
7. Berg, T., Belhumeur, P.N.: POOF: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In: CVPR (2013)
8. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3D human pose annotations. In: ICCV (2009), `http://www.eecs.berkeley.edu/~lbourdev/poselets`
9. Branson, S., Wah, C., Schroff, F., Babenko, B., Welinder, P., Perona, P., Belongie, S.: Visual recognition with humans in the loop. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 438–451. Springer, Heidelberg (2010)
10. Chai, Y., Lempitsky, V., Zisserman, A.: Symbiotic segmentation and part localization for fine-grained categorization. In: ICCV (2013)
11. Chai, Y., Rahtu, E., Lempitsky, V., Van Gool, L., Zisserman, A.: TriCoS: A tri-level class-discriminative co-segmentation method for image classification. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part I. LNCS, vol. 7572, pp. 794–807. Springer, Heidelberg (2012)
12. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)
13. Deng, J., Krause, J., Fei-Fei, L.: Fine-grained crowdsourcing for fine-grained recognition. In: CVPR (2013)
14. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: DeCAF: A deep convolutional activation feature for generic visual recognition. In: ICML (2014)
15. Duan, K., Parkh, D., Crandall, D., Grauman, K.: Discovering localized attributes for fine-grained recognition. In: CVPR (2012)
16. Farrell, R., Oza, O., Zhang, N., Morariu, V.I., Darrell, T., Davis, L.S.: Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance. In: ICCV (2011)
17. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE Transactions on Pattern Analysis and Machine Intelligence (2010)
18. Felzenszwalb, P.F., Huttenlocher, D.: Efficient matching of pictorial structure. In: CVPR (2000)

19. Fischler, M.A., Elschlager, R.A.: The representation and matching of pictorial structures. IEEE Transactions on Computers (January 1973), `http://dx.doi.org/10.1109/T-C.1973.223602`
20. Gavves, E., Fernando, B., Snoek, C., Smeulders, A., Tuytelaars, T.: Fine-grained categorization by alignments. In: ICCV (2013)
21. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR (2014)
22. ILSVRC: ImageNet Large-scale Visual Recognition Challenge (2010-2012), `http://www.image-net.org/challenges/LSVRC/2011/`
23. Jarrett, K., Kavukcuoglu, K., Ranzato, M., LeCun, Y.: What is the best multi-stage architecture for object recognition? In: ICCV (2009)
24. Jia, Y.: Caffe: An open source convolutional architecture for fast feature embedding (2013), `http://caffe.berkeleyvision.org/`
25. Khosla, A., Jayadevaprakash, N., Yao, B., Fei-Fei, L.: Novel dataset for fine-grained image categorization. In: FGVC Workshop, CVPR (2011)
26. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)
27. LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to hand-written zip code recognition. Neural Computation (1989)
28. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE, 2278–2324 (1998)
29. Liu, J., Belhumeur, P.N.: Bird part localization using exemplar-based models with enforced pose and subcategory consistency. In: ICCV (2013)
30. Liu, J., Kanazawa, A., Jacobs, D., Belhumeur, P.: Dog breed classification using part localization. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part I. LNCS, vol. 7572, pp. 172–185. Springer, Heidelberg (2012)
31. Maji, S., Kannala, J., Rahtu, E., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. Tech. rep. (2013)
32. Martinez-Munoz, G., Larios, N., Mortensen, E., Zhang, W., Yamamuro, A., Paasch, R., Payet, N., Lytle, D., Shapiro, L., Todorovic, S., Moldenke, A., Dietterich, T.: Dictionary-free categorization of very similar objects via stacked evidence trees. In: CVPR (2009)
33. Nilsback, M.E., Zisserman, A.: A visual vocabulary for flower classification. In: CVPR (2006)
34. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: ICVGIP (2008)
35. Parkhi, O.M., Vedaldi, A., Jawahar, C.V., Zisserman, A.: The truth about cats and dogs. In: ICCV (2011)
36. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.V.: Cats and dogs. In: CVPR (2012)
37. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: OverFeat: Integrated recognition, localization and detection using convolutional networks. CoRR abs/1312.6229 (2013)
38. Sfar, A.R., Boujemaa, N., Geman, D.: Vantage feature frames for fine-grained categorization. In: CVPR (2013)
39. Stark, M., Krause, J., Pepik, B., Meger, D., Little, J.J., Schiele, B., Koller, D.: Fine-grained categorization for 3D scene understanding. In: BMVC (2012)
40. Uijlings, J., van de Sande, K., Gevers, T., Smeulders, A.: Selective search for object recognition. IJCV (2013)

41. Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P.: Caltech-UCSD Birds 200. Tech. Rep. CNS-TR-2010-001, California Institute of Technology (2010)
42. Xie, L., Tian, Q., Hong, R., Yan, S., Zhang, B.: Hierarchical part matching for fine-grained visual categorization. In: ICCV (2013)
43. Yang, S., Bo, L., Wang, J., Shapiro, L.: Unsupervised template learning for fine-grained object recognition. In: NIPS (2012)
44. Yao, B., Bradski, G., Fei-Fei, L.: A codebook-free and annotation-free approach for fine-grained image categorization. In: CVPR (2012)
45. Yao, B., Khosla, A., Fei-Fei, L.: Combining randomization and discrimination for fine-grained image categorization. In: CVPR (2011)
46. Zhang, N., Farrell, R., Darrell, T.: Pose pooling kernels for sub-category recognition. In: CVPR (2012)
47. Zhang, N., Farrell, R., Iandola, F., Darrell, T.: Deformable part descriptors for fine-grained recognition and attribute prediction. In: ICCV (2013)
48. Zhang, N., Paluri, M., Ranzato, M., Darrell, T., Bourdev, L.: PANDA: Pose aligned networks for deep attribute modeling. In: CVPR (2014)