Symbiotic Segmentation and Part Localization for Fine-Grained Categorization

Yuning Chai
Dept. of Engineering Science
University of Oxford
chaiy@robots.ox.ac.uk

Victor Lempitsky
Skolkovo Institute of Science
and Technology (Skoltech)
lempitsky@skoltech.ru

Andrew Zisserman
Dept. of Engineering Science
University of Oxford
az@robots.ox.ac.uk

Abstract

We propose a new method for the task of fine-grained visual categorization. The method builds a model of the base-level category that can be fitted to images, producing high-quality foreground segmentation and mid-level part localizations. The model can be learnt from the typical datasets available for fine-grained categorization, where the only annotation provided is a loose bounding box around the instance (e.g. bird) in each image. Both segmentation and part localizations are then used to encode the image content into a highly-discriminative visual signature.

The model is symbiotic *in that part discovery/localization is helped by segmentation and, conversely, the segmentation is helped by the detection (e.g. part layout). Our model builds on top of the part-based object category detector of Felzenszwalb et al., and also on the powerful GrabCut segmentation algorithm of Rother et al., and adds a simple spatial saliency coupling between them. In our evaluation, the model improves the categorization accuracy over the state-of-the-art. It also improves over what can be achieved with an analogous system that runs segmentation and part-localization independently.*

1. Introduction

Fine-grained visual categorization is the task of distinguishing between sub-ordinate categories, e.g. between "tree sparrow", "Ivory gull" and "Anna hummingbird", which all belong to the base level category "bird". Several recent works have pointed out two aspects, which distinguish visual categorization at the subordinate level from that at the base level.

First, in subordinate classification it often happens that two similar classes can only be distinguished by the appearance of localized and very subtle details (such as the color of the beak for bird classes or the shape of the petal edges for flower classes). With generic classification approaches these fine differences often get "swamped" by the bulk of the image, whenever encoding of the image content into a visual signature of some sort is performed. Therefore, [5, 24, 32, 34, 35] focused on the localization of these discriminative image parts as a precursor to categorization. Once the discriminative parts are localized, they are encoded into separate parts of the visual signature, enabling the classifier to pick up on the fine differences in those parts.

The second distinguishing aspect is the role of the background. It is well known [13] that at the base category level the background often provides valuable context for categorization. However, [10, 22, 24] demonstrated that at the subordinate category level, the background is seldom discriminative and it is beneficial to segment out the foreground and to discard the visual information in the background. [10] further demonstrated that increasing the accuracy of foreground segmentation at training time directly translates into an increase in accuracy of subordinate-level categorization at test time.

In the light of all this evidence, it is natural to investigate the combination of part localization and foreground segmentation for fine-grained categorization, and their interaction in combination is the topic of this work. Our least surprising finding (which nevertheless translates into a very competitive categorization system) is that a simple concatenation of visual signatures, provided by a system that performs part localization and by a system that performs foreground segmentation, leads to improved categorization accuracy (as compared to classifiers operating with each of the two signatures individually).

More interestingly, we demonstrate that the accuracy of fine-grained categorization can be further boosted if part localization and foreground segmentation are performed together, so that the outcomes of both processes aid each other. As a result, better segmentation can be obtained by taking into account part localizations, and, likewise, more semantically meaningful and discriminative parts can be learned and localized if foreground masks are taken into account. We implement this feedback loop via the energy minimization of a joint functional that incorporates the consistency between part localization and foreground segmentation as one of the terms. The resulting *symbiotic* system achieves a better categorization performance compared to the system obtained by a mere concatenation of two visual

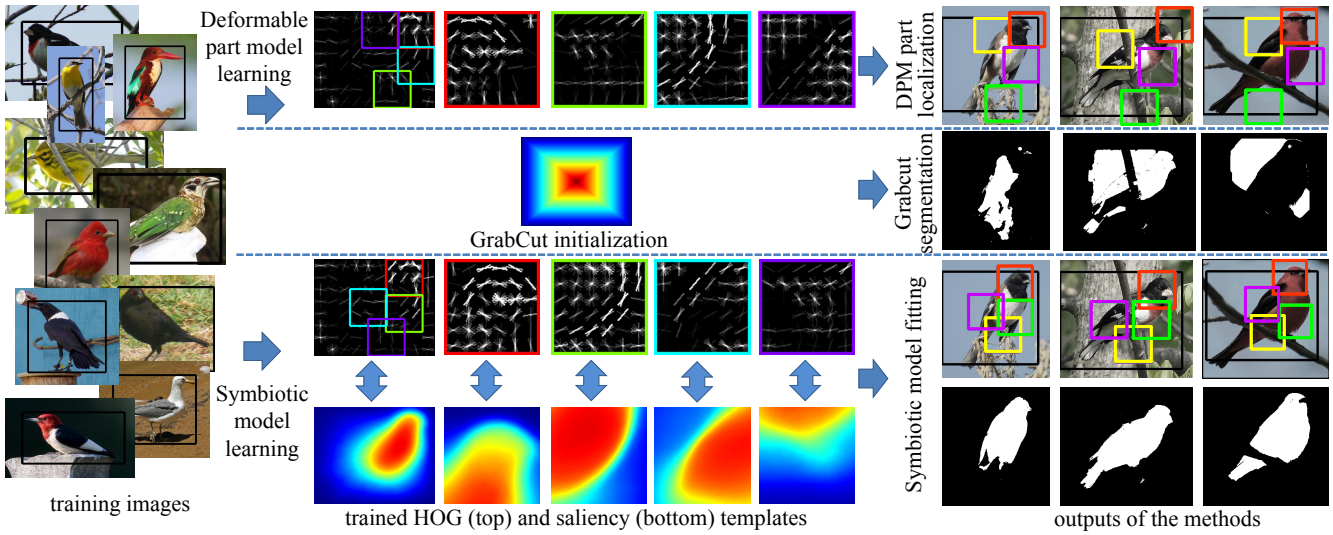

Figure 1. Best viewed in color. We demonstrate our system using images from the Caltech-UCSD Bird dataset. Left: examples of the training images. Black frames indicate the provided ground truth bounding box. **Top**: a stand alone Deformable Part Model (DPM) with its results to the right. **Center**: GrabCut automatically segments the images using the outside of the given bounding box as background and a prior foreground saliency map for the region inside the bounding box. **Bottom**: our approach, which trains a symbiotic set of detector templates and saliency maps and applies them jointly to images. As a result it achieves a considerable improvement in segmentation accuracy, part-localization consistency, and the ultimate goal of fine-grained classification accuracy. (The saturation in the output images is reduced for illustration).

signatures (discussed above). Overall, our symbiotic system outperforms the previous state-of-the-art on all datasets considered in our experiments (both the 2010 and 2011 version of Caltech-UCSD Birds, and Stanford Dogs). This symbiotic system is the main contribution of the paper.

As a coda, we investigate the gains in performance by using additional annotation, and show that although training performance is near saturation, significant improvements are still possible at test time; thus confirming similar findings (e.g. a human in-the-loop [8]) in recent literature.

2. Related Work

There is a line of work stretching back over a decade on the interplay between segmentation and detection. In early works, object category detectors simply proposed foreground masks [4, 18]. Later methods used these masks to initialize graph-cuts based segmentations [7] that could take advantage of image specific color distributions, giving crisper and more accurate foreground segmentations [17, 19, 26].

In the *poselet* line of research [6] the detectors are for parts, rather than for entire categories, but again the poselet-detectors can predict foreground masks for object category detection and segmentation [9, 20]. Whether the parts arise from poselets [35] or are discovered from random initializations [33], there are benefits in comparing objects in fine-grained visual categorization tasks at the part level where subtle discriminative features are more evident. We demonstrate, however, that the parts discovered in the absence of supervision are less discriminative than those discovered with the help of the segmentation process as is done in our method.

Co-segmentation methods have been successful in building nuanced models of a base-level class in an unsupervised way. A representative early work in this area is LOCUS [31]. The more recent methods such as [10] used cosegmentation-based models for fine-grained categorization. These methods however do not attempt to model mid-level discriminative parts.

The closest work to ours is that of [32]. It also accomplishes unsupervised learning of a deformable part model in order to find discriminative parts for fine-grained categorization. An earlier method had used the image as a bounding box for learning a deformable parts model for scene classification [23]. Again, neither of these use segmentation to aid the part learning and localization.

In summary, although the synergy between segmentation and detection has long been recognized [16], the interplay between part localization and segmentation has not been investigated in the context of fine-grained categorization (to the best of our knowledge). By exploiting this interplay, the proposed approach is able to achieve a significant improvement in the categorization accuracy.

3. Symbiotic Segmentation and Localization

We start with an overview of the system. It is built around a model of the base category (e.g. bird) which includes a deformable part model $\mathcal{W}$ and a set $\mathcal{S}$ of saliency

maps each associated with a part or root of the DPM. At test time, given a pre-trained model, the model is fitted to an image I via the minimization of the following three-term cost function:

$$E(\mathbf{p},\mathbf{f},\mathbf{c}|\mathcal{W},\mathcal{S},I) = \alpha E^{DPM}(\mathbf{p}|\mathcal{W},I) + \beta E^{GC}(\mathbf{f},\mathbf{c}|I) + E^{C}(\mathbf{p},\mathbf{f}|\mathcal{S}) \quad (1)$$

Here, the minimization is performed over the part localizations $\mathbf{p}$, the foreground mask $\mathbf{f}$, and the color distributions of the foreground and background $\mathbf{c}$. α and β are weights controlling the balance between the energy terms. The recovered part localizations $\mathbf{p}$ and the foreground segmentation $\mathbf{f}$ are then used to encode the image content into a highly-discriminative visual signature as discussed in the next section. The model is intuitive: the first two mutually independent terms in (1) correspond to the popular models we build upon, E^{DPM} denotes a Deformable Part Model (DPM) [14] energy; while E^{GC} denotes a GrabCut [27] energy. With the introduction of a third (consistency) energy term E^{C} that takes a pre-trained saliency model $\mathcal{S}$ we penalize the cases where the foreground segmentation $\mathbf{f}$ and the part locations $\mathbf{p}$ do not agree. We postpone its definition to Sec. 3.1 and first discuss the variables in (1) in more detail.

Deformable part model $\mathcal{W} = \{\mathbf{w}_t\}$: here, we use a multi-component Deformable Part Model (DPM) [14] consisting of several mixtures of parts, where each part is described by a HOG template and a geometric location prior. We denote the number of mixture components N, and the number of parts in each component M. We omit extra indices for different mixture components and use $\mathbf{w}_0$ to describe the root HOG template for each component. $\mathbf{w}_t$ then denotes the parameters of the t-th part (the HOG template and the geometric prior).

Saliency model $\mathcal{S} = \{\mathbf{s}_t\}$: we associate with the root and each part $\mathbf{w}_t$ of the deformable part model an extra map $\mathbf{s}_t$ that indicates the foreground probability. Pixels of this saliency map thus have values between -1 and 1, with 1 indicating a high chance of the pixel being foreground and -1 otherwise. An example of a set of saliency maps is shown in the center of the bottom row of Fig. 1.

Part localizations $\mathbf{p} = \{\mathbf{p}_t\}$: this variable denotes the location (the bounding box coordinates) of all detected parts in an image. Only one mixture component is active for a single image. The localization of a particular part template $\mathbf{w}_t$ is denoted $\mathbf{p}_t$. The part localizations are shown as colored bounding boxes in the output images of Fig. 1.

Color distributions $\mathbf{c} = \{\mathbf{c}_{-1}, \mathbf{c}_1\}$: following GrabCut [27], we model the distribution of colors in the image in the foreground and the background as Gaussian mixtures in RGB space (denoted $\mathbf{c}_1$ and $\mathbf{c}_{-1}$ respectively).

Foreground segmentation $\mathbf{f}$: this map assigns each pixel the value 1 if it is foreground, and -1 if it is background. Examples of the binary segmentations are shown as binary maps in Fig. 1.

Note that $\mathbf{p}$, $\mathbf{f}$, $\mathbf{c}$ are specific to an image I, while $\mathcal{W}$ and $\mathcal{S}$ are global parameters describing the base-level category (e.g. bird or dog). These parameters can be learned from a dataset $\mathcal{I}$ of images containing instances of this base category as discussed in Sec. 3.2.

3.1. Optimization

We begin by describing the consistency term in (1), and then detail the minimization of the entire cost function.

Consistency term: E^{C}: this is defined as the sum of a set of distances (or equivalently as as a sum of correlations):

$$E^{C}(\mathbf{p},\mathbf{f}|\mathcal{S}) = \frac{1}{2}\sum_t ||\mathbf{m}_t(\mathbf{p}_t,\mathbf{f}) - \mathbf{s}_t||_2^2 \quad (2)$$

$$= \frac{1}{2}\sum_t ||\mathbf{m}_t(\mathbf{p}_t,\mathbf{f})||_2^2 - 2\langle \mathbf{m}_t(\mathbf{p}_t,\mathbf{f}), \mathbf{s}_t\rangle + ||\mathbf{s}_t||_2^2$$

$$= -\sum_t \langle \mathbf{m}_t(\mathbf{p}_t,\mathbf{f}), \mathbf{s}_t\rangle + C \quad (3)$$

where $\mathbf{m}_t(\mathbf{p}_t,\mathbf{f})$ is a binary map $\{-1,1\}$ clipped from the segmentation mask $\mathbf{f}$ by the localized part bounding box $\mathbf{p}_t$. This map is resized to the size of a saliency map $\mathbf{s}_t$, which is denoted as θ_t. C is a constant with respect to $\mathbf{p}_t$ and $\mathbf{f}$ and therefore can be ignored during the optimization. $||\mathbf{m}_t(\mathbf{p}_t,\mathbf{f})||_2^2$ is constant for the reason that $\mathbf{m}_t$ only contains pixel values of either -1 or 1 and hence the squared norm is simply the number of pixels specified by the size θ_t, and does not depend on $\mathbf{p}_t$ and $\mathbf{f}$.

We optimize the cost function (1) using a block-coordinate-descent pattern, that is, alternating between updating part localizations $\mathbf{p}$ while fixing the foreground segmentation $\mathbf{f}$ and color $\mathbf{c}$, and vice versa.

Updating part localizations $\mathbf{p}$. When finding the best part localization $\mathbf{p}$ (given the DPM $\mathcal{W}$, the saliency model $\mathcal{S}$ and the foreground segmentation $\mathbf{f}$), E^{GC} can be ignored and we are left with the original DPM term and the consistency term:

$$\min_{\mathbf{p}} \alpha E^{DPM}(\mathbf{p}|\mathcal{W},I) + E^{C}(\mathbf{p},\mathbf{f}|\mathcal{S}) \quad (4)$$

We modified the standard off-the-shelf DPM detector [14] to solve (4). The DPM energy E^{DPM} from [14] can be written as:

$$E^{DPM}(\mathbf{p}|\mathcal{W},I) = -R(\mathbf{p}_0,\mathbf{w}_t) - \sum_{t\neq 0} D(\mathbf{p}_t,\mathbf{w}_t,\mathbf{p}_0) \quad (5)$$

$$D(\mathbf{p}_t,\mathbf{w}_t,\mathbf{p}_0) = R(\mathbf{p}_t,\mathbf{w}_t) + Q_t(\mathbf{p}_t,\mathbf{p}_0) \quad (6)$$

$R(\mathbf{p}_t,\mathbf{w}_t)$ is the HOG-template filter response map of the t-th root or part template. Q_t is a quadratic function of the

relative location of the part and the root that penalizes the atypical geometric configurations.

Minimization of (4) is then equivalent to the minimization of (5) with the following modification of the response function $R(\mathbf{p}, \mathcal{W}) \rightarrow R'(\mathbf{p}, \mathbf{f}, \mathcal{W}, \mathcal{S})$:

$$R'(\mathbf{p}_t, \mathbf{f}, \mathbf{w}_t, \mathbf{s}_t) = \alpha R(\mathbf{p}_t, \mathbf{w}_t) + \mathbf{m}_t(\mathbf{p}_t, \mathbf{f}) \otimes \mathbf{s}_t \quad (7)$$

Here, $\otimes$ is the convolution operator and α is a scalar constant which balances between the two information sources. The modified response function is then passed to an off-the-shelf DPM solver which finds the optimum $\mathbf{p}$ for (4) via tree dynamic programming.

Updating foreground segmentation $\mathbf{f}$ and color models $\mathbf{c}$. Assuming that part localizations $\mathbf{p}$ are fixed, the minimization

$$\min_{\mathbf{f}} \beta E^{GC}(\mathbf{f}, \mathbf{c}|I) + E^{C}(\mathbf{p}, \mathbf{f}|\mathcal{S}) \quad (8)$$

can be accomplished with an appropriately modified GrabCut algorithm.

Recall that GrabCut alternates the color model updates and the segmentation updates. Since the consistency term (2) does not depend on color model $\mathbf{c}$, the color model update step is left unchanged compared to the original GrabCut [27]. Let us now focus on the foreground segmentation update (given part localizations $\mathbf{p}$ and the color model $\mathbf{c}$).

Recall that this update within the original GrabCut minimizes the following energy:

$$E^{GC}(\mathbf{f}, \mathbf{c}|I) = \sum_x U_x + \sigma \sum_{(x,x')} V_{x,x'} \quad (9)$$

$$U_x = \mathbf{f}(x)\{U^{\text{GMM}}_{-1}(I(x)) - U^{\text{GMM}}_{1}(I(x))\} \quad (10)$$

$$V_{x,x'} = |\mathbf{f}(x) - \mathbf{f}(x')|\ v\,(I(x) - I(x')) \quad (11)$$

$I(x)$ denotes an RGB value at pixel x, (x, x') spans all pairs of adjacent pixels, v is the binary Ising potential weighted according to the contrast observed between the two pixels. The unary potential U^{GMM}_k is equal to the log-likelihood of $I(x)$ under the Gaussian mixture $\mathbf{c}_k(I(x))$, where k is the foreground/background label $\{-1, 1\}$ of pixel x.

To add the consistency term (2), we first re-express it using image pixel-based terms:

$$\begin{aligned} E^{C}(\mathbf{p}, \mathbf{f}|\mathcal{S}) &= \tfrac{1}{2} \textstyle\sum_x \sum_t \frac{1}{r_t(\mathbf{p}_t)} (\mathbf{n}_x(\mathbf{p}_t, \mathbf{s}_t) - \mathbf{f}(x))^2 \\ &= - \textstyle\sum_x \mathbf{f}(x) \sum_t \frac{1}{r_t(\mathbf{p}_t)} \mathbf{n}_x(\mathbf{p}_t, \mathbf{s}_t) + C \end{aligned} \quad (12)$$

x describes pixel location, and $\mathbf{f}(x)$ denotes the binary foreground-background label at position x. $\mathbf{n}(\mathbf{p}_t, \mathbf{s}_t)$ describes a real valued saliency map of the same size as the input image. It has all pixel values equal to 0 except for the window specified by $\mathbf{p}_t$, which is filled with an appropriately resized $\mathbf{s}_t$. $\mathbf{n}_x$ is then the value of $\mathbf{n}$ at location x. Note that to ensure the equivalence of (12) and (2), each term in (12) is reweighted by the reciprocal of the $r_t(\mathbf{p}_t)$, which is the ratio between the number of pixels in $\mathbf{s}_t$ specified by the size hyper-parameter θ_t and the number of pixels defined in the window in $\mathbf{p}_t$. The squared terms from expanding (12) do not depend on $\mathbf{p}$ and $\mathbf{f}$ for the same reason as in (3).

Adding (12) into (9) keeps the pairwise terms unchanged, while modifying the unary potential, $U_x \rightarrow U'_x$:

$$U'_x = \beta U_x - \sum_x \mathbf{f}(x) \sum_t \frac{1}{r_t(\mathbf{p}_t)} \mathbf{n}_x(\mathbf{p}_t, \mathbf{s}_t) \quad (13)$$

The modified energy can still be minimized exactly via graph cut.

In conclusion, the minimization of (1) alternates between three steps: **(a)** optimizing for $\mathbf{p}$ with the help of a DPM solver with modified filter responses according to (7), **(b)** estimating the color model $\mathbf{c}$ (standard GMM estimation step within GrabCut) and **(c)** optimizing for $\mathbf{f}$ using GrabCut with modified unary energy as defined in (13).

3.2. Learning the Model

The DPM model $\mathcal{W}$ and the saliency model $\mathcal{S}$ are trained using a set $\mathcal{I}$ of training images. We learn the model progressively, starting with the HOG-filters and saliency mask corresponding to the root, and then proceeding to the parts.

Learning the root parameters. We start with the training of the HOG template for the root filter $\mathbf{w}_0$ of the DPM model. For the most part we follow the approach of Felzenszwalb *et al.* (c.f. section 5.2 in [14]). Thus, the HOG templates for root filters are in the mixture components via latent SVM training (we use a separate unrelated dataset as a source of negative examples; and constrain the root filters to overlap with user-provided boxes by at least 70%). At the same time, we run GrabCut on all training examples (using bounding box annotations), and estimate the root saliency map $\mathbf{s}_0$ corresponding to root filters by averaging the segmentation masks (as detailed below).

Discriminative part discovery. We then use a standard DPM approach to discover repeatable parts $\mathbf{w}_{t,\forall t \neq 0}$ with an important modification. In [14], "interesting" parts are discovered greedily (as discussed in [14]) by covering the high-energy (large gradient magnitude) parts of the root HOG-template. In our case, we modify this interestingness measure by multiplying the HOG magnitude by the root saliency maps estimated for each component. In this way, we constrain the discovery process to parts which overlap substantially with the foreground (as estimated by a GrabCut). We found this modification to be important to make the learned parts consistent with our model (1), but also to discover more semantically meaningful parts. We come back to the issue of unsupervised part discovery in the experiments section. After the discovery, we proceed with the

	[32]	[35]	[1]	[2]	Symb*	Symb
Birds11	-	28.2	-	56.8	56.6	**59.4**
Birds10	28.2	-	30.2	-	46.5	**47.3**
Dogs	38.0	-	-	-	44.1	**45.6**

Table 1. Performance on the three fine-grained categorization datasets. Symbotic system ("Symb") consistently outperforms previously published results. "Symb*" shows the accuracy of the system, for which classifiers were trained on the sets not augmented by left-right mirroring. The authors of [35] have confirmed to have used mA rather than mAP in their paper.

standard DPM training, and fit the learnt DPM to each training image.

Learning the saliency model $\mathcal{S}$. Given the part localizations and the GrabCut segmentations of all training images, we set the saliency mask for each part to be the pixel-wise mean of all segmentation masks cutouts, corresponding to the locations of this part (i.e. $\mathbf{s}_t = \frac{1}{|\mathcal{I}|}\sum_{I\in\mathcal{I}} \mathbf{m}_t(\mathbf{p}_t^I, \mathbf{f}^I)$).

4. Experimental Results

The empirical evaluation is carried out on three benchmark datasets for fine-grained image classification – the Caltech-UCSD Birds 2010 and 2011, and Stanford Dogs. Both versions of the Caltech-UCSD Birds [30] contain 200 bird categories. While the 2010 version only has 15 training and around 15 test images per class, the 2011 version increased both numbers to 30. Evaluations are on both the 2010 and 2011 versions of the Caltech Birds, in order to compare to as many state-of-the-art works as possible. The Stanford Dogs dataset [15] consists of 120 dogs species and has around 100 training images/70 test images per class. The images are a carefully filtered subset of ImageNet.

In all experiments, we make use of the provided bounding boxes around the object during both training and testing, as do most of the approaches we compare to. During preprocessing, all images are first resized such that the bounding box has the longest dimension equal to 300 pixels. Images are cropped to include the bounding box together with a maximum 50 pixel wide strip around the box. This is important for any GrabCut-related steps as the background can be better estimated using the strip. Each dataset is augmented with the left-right mirrored versions of its training images, as this typically yields a 3-5% improvement over not doing so (for reference we also give final results without such mirroring).

The symbiotic model is fitted to every train and test images using 5 alternation iterations (the convergence is observed after 3 iterations in most cases). It takes about 10 seconds to fit the model to a typical image. The parameters α and β were set to 0.1 and 4 respectively (we find the final accuracy to be not too sensitive to the variation of these parameters). The choice of the parameters M, N is discussed below.

Classification Process. The symbiotic model outputs one binary segmentation and a set of detected part bounding boxes for a given image. Descriptors are extracted from each of them individually, i.e., one feature vector, $\mathbf{x}^{SEG}$, for the foreground region in the segmentation, and a feature vector for each of the parts apart from the root template. A feature vector is not included for the root template as it would be too redundant with $\mathbf{x}^{SEG}$. We denote the concatenation of all part features as $\mathbf{x}^{PART}$. If the final feature dimension is D, we use $D/2$ for $\mathbf{x}^{PART}$ and the other $D/2$ for $\mathbf{x}^{SEG}$.

Each region (i.e. the foreground and the box of each part) is encoded by: (1) LLC-encoded [29] Lab color histogram vector, and (2) Fisher vector [25] aggregating SIFT features (the implementation [11] was adopted). Both features are ℓ_2 normalized after encoding and then concatenated. Finally, after another $\ell 2$ normalization, $\mathbf{x}^{SEG}$ and $\mathbf{x}^{PART}$ are concatenated. A conventional multi-class 1-vs-rest linear support vector machine (SVM) is used for the final fine-grained classification (the regularization strength is set by cross-validation).

To encode the foreground, we use a k-means Lab vocabulary of size 512, and a SIFT GMM with 128 components. The resulting feature vector $\mathbf{x}^{SEG}$ is has 20992 dimensions. When encoding parts, we choose the size of the vocabulary so that $\mathbf{x}^{PART}$ and $\mathbf{x}^{SEG}$ are always the same length (i.e. 20992 dims each), no matter how many parts and mixture components are used.

Performance Measures. We evaluate the categorization performance of several baselines and variations of our system, and report two performance measures for this: (1) **Mean accuracy (mA)**: for each class we measure the proportion of test images of the class that are classified correctly (as belonging to this class). The proportion is then averaged over all classes. This measure is the one used in most previous works. (2) **Mean average precision (mAP)**: For each class, we evaluate the SVM score of the class' classifier for the entire dataset. Once the dataset is ordered by decreasing score, the average precision (AP) of the returned list is computed (i.e. the area under the precision-recall curve). The AP numbers are averaged over all classes. This measure is more relevant than mean accuracy (mA) for some applications (e.g. Web image search).

4.1. Results and Comparisons

Overall, our complete system surpasses all previously published results on all three datasets (Tab. 1). The models learned by the symbiotic system can be seen in Fig. 1 and Fig. 2. The relative importance of the model components, as well as the net effect of the "symbiosis" between the segmentation and part localization, are evaluated in Tab. 2.

In the table, we compare the categorization accuracy of the systems resulting from applying GrabCut alone or DPM

ID	Model fitting	Descriptor	Birds11		Birds10		Dogs	
			mA	mAP	mA	mAP	mA	mAP
1	taking whole bounding box	$\mathbf{x}^{SEG}$	40.7	32.5	27.9	20.0	39.7	33.0
2	GrabCut segmentation	$\mathbf{x}^{SEG}$	51.1	40.4	39.3	26.7	42.2	33.9
3	Symbiotic model fitting	$\mathbf{x}^{SEG}$	57.5	41.9	42.1	25.2	47.3	37.8
4	DPM part localization	$\mathbf{x}^{PART}$	38.6	27.3	26.7	15.1	22.2	17.0
5	Symbiotic model fitting	$\mathbf{x}^{PART}$	52.0	36.0	40.1	23.6	34.8	28.5
6	GrabCut + DPM (independent)	$[\mathbf{x}^{SEG}; \mathbf{x}^{PART}]$	54.4	46.6	41.7	30.4	41.3	35.8
7	Symbiotic model fitting	$[\mathbf{x}^{SEG}; \mathbf{x}^{PART}]$	**59.4**	**52.1**	**47.3**	**35.4**	**45.6**	**40.7**

Table 2. A detailed comparison with baselines (no model fitting, segmentation only, part localization only). Note that segmentations produced by the symbiotic model allow for more discriminative signatures than those produced with GrabCut alone (#3 vs. #2), while parts learned and localized by the symbiotic model are more discriminative than those learned and localized by DPM (#5 vs. #4). Finally, categorization with full signatures produced by symbiotic model is better than categorization based on the concatenation of segmentation-based and part-based signatures produced by GrabCut and DPM run independently (#7 vs #6). All these improvements are due to the fact that part localization and segmentation processes assist each other within the proposed symbiotic model.

part localization alone, while keeping the rest of parameters (initialization, feature encoding, etc.) fixed. Notably, a considerable improvement over a GrabCut-based system (line 2) is observed even if we only use the segmentation-based descriptor $\mathbf{x}^{SEG}$ in our system (line 3), thus highlighting that segmentations obtained by our systems are better (at least for further categorization). Likewise, the same improvement is observed for part localization, when the segmentation process is used to aid part discovery and fitting, as opposed to using a DPM model on its own (line 5 vs line 4). Finally, and most importantly, the symbiotic system improves considerably in all measures on all three datasets when compared to the system that gets the same visual signature by running the classification and the part localization processes independently and concatenating the corresponding signatures (line 7 vs line 6).

The interaction between the segmentation and the part localization processes are further shown in Fig. 3 and Fig. 4. Note, that in the case of Fig. 3, we used the same deformable part model $\mathcal{W}$ (learned within the symbiotic model) but evaluated it with and without the help of the segmentation process. In Fig. 4, we simply compare the segmentations obtained by our system and by GrabCut. In both cases, it can be seen how symbiosis between the part localization and the segmentation improve the performance of each process.

We note that the improvement over the baselines (especially over the GrabCut baseline) is smaller for the Dogs dataset rather than for the Birds datasets. We attribute this fact to a greater pose variability for dogs that is harder to cope with for the deformable parts model. At the same time, dogs have a nice roundish shape which makes them very appropriate for GrabCut (so that the aid from the parts localization is not needed in most cases). The performance of the DPM on dogs can be potentially improved by having more mixture components. However, as discussed below, it might hurt the generalization in the categorization step, and especially since we keep the feature dimension of $\mathbf{x}^{PART}$ the same. Post-processing as suggested in [35], may also be useful in this case.

Influence of the parameters. We have further evaluated the influence of the size of the deformable parts model on the categorization accuracy, namely N (number of mixture components) and M (the number of parts per component). As discussed in [14], in the context of detection a larger N increases the non-linearity of the model while also increasing data fragmentation. Meanwhile, an M has to strike a balance between having too many parts some of which are not detectable and having too few parts, which will make the detector less powerful.

In the context of building the base-class model for fine grained classification, M and N have some additional meaning. While large N may also increase the data fragmentation within some subordinate classes, potentially having large N may also attribute different subordinate classes to different components, thus making the categorization easier. At the same time, picking the value for M faces the usual choice between feature repeatability and the discriminating power. The more parts the model has, the more discriminative information it can provide into $\mathbf{x}^{PART}$. However, it becomes more difficult to detect parts repeatedly at the same semantic "locations".

We mainly selected these 2 parameters based on visual feedback during the training stage. But we also did some quantitative evaluation using different settings for the Bird 2011 dataset, as shown in Tab. 3. Overall, for the bird datasets, we chose $N = 1$ and $M = 4$, while $N = 2$ and $M = 4$ seems to be more reasonable for the dogs dataset[1].

4.2. Experiments with Extra Annotation

Looking at Tab. 2, one can notice that generally the segmentation-based signatures outperform part-localization

[1]We use the functionality of the code [14] which allows to apply each DPM mixture component twice (once with mirroring and once without) during training and test.

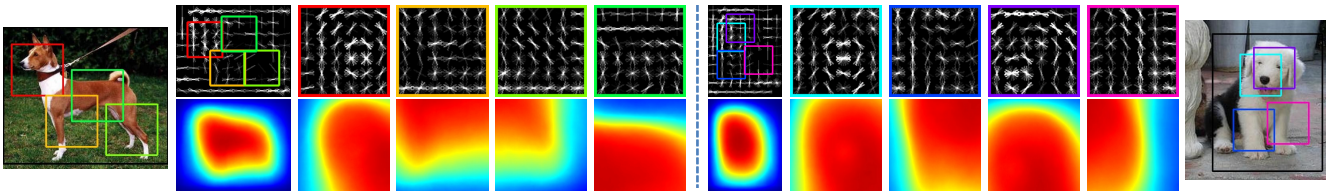

Figure 2. Trained $\mathcal{W}$ and $\mathcal{S}$ for the dogs dataset. After learning a symbiotic model, the two mixture components (shown side-by-side) happen to correspond to a more profile and a more frontal views.

$N \times M$	1×8	1×4	1×2	1×1	2×4	2×2	4×2
mA	59.2	**59.4**	58.2	58.3	57.6	55.9	52.9
mAP	**54.3**	52.1	49.2	45.9	52.0	47.2	46.1

Table 3. Effect of different choice of N and M evaluated on the Caltech-UCSD Birds 2011. The loss in accuracy with higher number of mixture components indicates that the complexity of a bird pose does not justify more than one mixture component in our model.

based signatures very considerably. Only by combining segmentation and part localization (lines 6 and 7 in the table) we can see a consistent benefit from having part localization in the system. One natural question is whether the performance of part localization is inherently limited or is this a problem with segmentation-supervised and, particularly, unsupervised part discovery?

To address this question we used the extensive annotations available for the Birds 2011. Apart from the bounding boxes, there are 15 part locations annotated per image. These parts include, e.g. beak, eyes, feet, etc. Given these annotations, we evaluated what would be achievable if we move away from unsupervised parts discovery and localization to supervised parts learning, or even using supervised parts localization during both training and testing (the latter would correspond to the scenario of asking the user to annotate some parts in the test image, thus approaching the human-in-the-loop approach investigated in [8]).

For simplicity, we considered a single part – a head of a bird, which leads to a setup that is similar to [24]. Thus, we first made use of the annotated head locations and trained a head detector (which was a mixture of HOG templates). This detector was used to locate heads in bird images. The first two experiments in Tab. 4 correspond to this setup. In a second set of experiments, we used the ground truth (rather than detected) head locations at all stages. Through these batch of experiments we followed the rest of our pipeline (i.e. extracting feature from parts/foreground segmentation and concatenating them, etc.).

As shown in Tab. 4, the resulting systems were able to surpass the performance of the symbiotic system even when only using the trained head detector. Using ground truth head localizations, the gap in the achieved accuracy compared to the symbiotic system (and, naturally, all other systems evaluated on this task) becomes very large. Overall, our conclusion here is that part localization has

localization	Descriptor	GT	mA	mAP
det. head	$\mathbf{x}^{PART}$	trn	52.4	31.9
GC + det. head	$[\mathbf{x}^{SEG}; \mathbf{x}^{PART}]$	trn	61.0	51.2
GT head	$\mathbf{x}^{PART}$	trn/tst	60.2	45.5
GC + GT head	$[\mathbf{x}^{SEG}; \mathbf{x}^{PART}]$	trn/tst	69.5	62.2

Table 4. Using extra annotation on Caltech-UCSD Birds 2011. The top two rows show the results if the head detector is trained using human annotation rather than unsupervisedly trained, while the bottom rows show the accuracies if the head position is given even during test time.

a great potential for fine-grained categorization. While the segmentation-based discovery and localization that we present in this paper is a definite step forward, compared to fully unsupervised part discovery and localization, there is still a big room for improvement to unleash the full potential of part localization for base-class modeling.

5. Conclusion

We have introduced and demonstrated the worth of a symbiotic part localization and segmentation model for fine-grained categorization. It successfully pulls together a number of recent research strands: the use of distinctive parts for registration when discriminating sub-ordinate categories [5, 24, 32, 34, 35]; unsupervised discovery of mid-level discriminative patches [23, 28, 32]; learning a DPM given only weak annotation (a loose bounding box compared to the tight boxes provided in PASCAL VOC) [3, 12, 21]; and, improving segmentations using a lite spatial model [31].

It also opens up new research questions: how can the model be extended from loose bounding box annotation to (even weaker) image level annotation? How should the number of components and parts be determined automatically? How should humans be used in-the-loop [8] to provide annotation at test time (based on the results from section 4.2)?

Acknowledgements. Financial support was provided by ERC grant VisRec no. 228180.

References

[1] A. Angelova and S. Zhu. Efficient object detection and segmentation for fine-grained recognition. In *CVPR*, 2013.

[2] T. Berg and P. N. Belhumeur. POOF: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In *CVPR*, 2013.

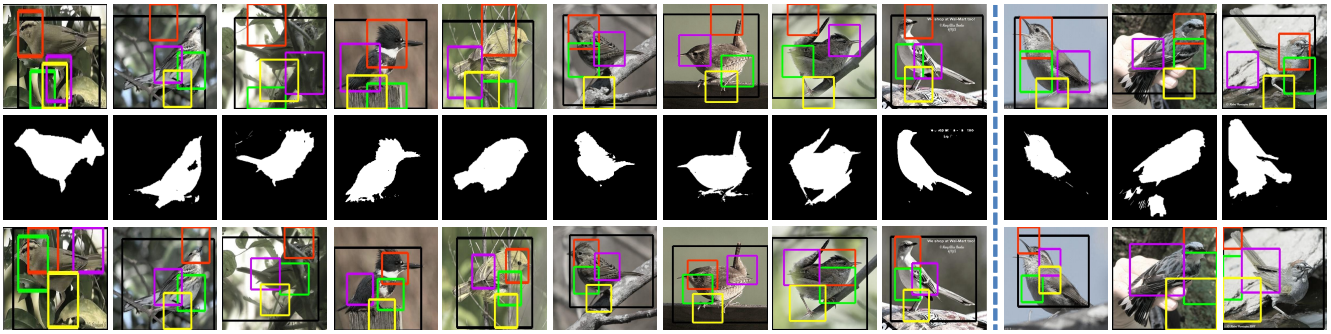

Figure 3. Examples taken from the Caltech Birds dataset. Top: part localizations using the symbiotically trained DPM, but fitted without the guidance of segmentation. Bottom: the same DPM model fitted with the help of segmentation (i.e. our full system). The segmentations are shown in the middle. The last three columns show some failure cases where segmentations hurts part localization.

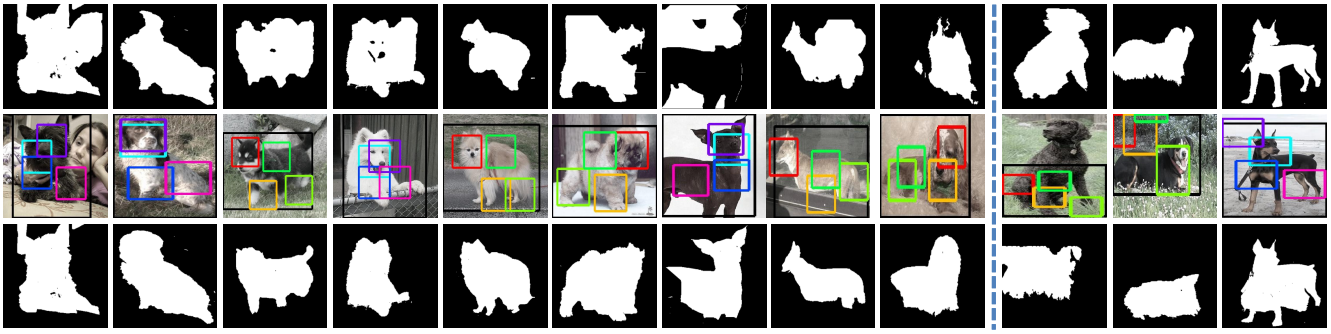

Figure 4. Examples taken from the Stanford Dogs dataset. Top: stand-alone segmentation results using GrabCut. Bottom: segmentation results with the help from the localized parts shown in the middle row (our full system). The last three columns show sample failure cases.

[3] M. B. Blaschko, A. Vedaldi, and A. Zisserman. Simultaneous object detection and ranking with weak supervision. In *NIPS*, 2010.

[4] E. Borenstein and S. Ullman. Class-specific, top-down segmentation. In *ECCV*, 2002.

[5] L. D. Bourdev, S. Maji, and J. Malik. Describing people: A poselet-based approach to attribute classification. In *ICCV*, 2011.

[6] L. D. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009.

[7] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *PAMI*, 2004.

[8] S. Branson, C. Wah, F. Schroff, B. Babenko, P. Welinder, P. Perona, and S. Belongie. Visual recognition with humans in the loop. In *ECCV*, 2010.

[9] T. Brox, L. D. Bourdev, S. Maji, and J. Malik. Object segmentation by alignment of poselet activations to image contours. In *CVPR*, 2011.

[10] Y. Chai, E. Rahtu, V. Lempitsky, L. V. Gool, and A. Zisserman. Tricos: A tri-level class-discriminative co-segmentation method for image classification. In *ECCV*, 2012.

[11] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *BMVC*, 2011.

[12] O. Chum and A. Zisserman. An exemplar model for learning object classes. In *CVPR*, 2007.

[13] S. K. Divvala, D. Hoiem, J. Hays, A. A. Efros, and M. Hebert. An empirical study of context in object detection. In *CVPR*, 2009.

[14] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 2010.

[15] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, CVPR*, 2011.

[16] I. Kokkinos and P. Maragos. Synergy between object recognition and image segmentation using the expectation-maximization algorithm. *PAMI*, 2009.

[17] M. P. Kumar, P. H. S. Torr, and A. Zisserman. OBJ CUT. In *CVPR*, 2005.

[18] B. Leibe and B. Schiele. Interleaved object categorization and segmentation. In *BMVC*, 2003.

[19] A. Levin and Y. Weiss. Learning to combine bottom-up and top-down segmentation. In *ECCV*, 2006.

[20] M. Maire, S. X. Yu, and P. Perona. Object detection and segmentation from joint embedding of parts and pixels. In *ICCV*, 2011.

[21] M. H. Nguyen, L. Torresani, F. de la Torre, and C. Rother. Weakly supervised discriminative localization and classification: a joint learning process. In *ICCV*, 2009.

[22] M. E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008.

[23] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *ICCV*, 2011.

[24] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, 2012.

[25] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, 2010.

[26] D. Ramanan. Using segmentation to verify object hypotheses. In *CVPR*, 2007.

[27] C. Rother, V. Kolmogorov, and A. Blake. "grabcut": interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3), 2004.

[28] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012.

[29] J. Wang, J. Yang, K. Yu, F. Lv, T. S. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010.

[30] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[31] J. M. Winn and N. Jojic. Locus: Learning object classes with unsupervised segmentation. In *ICCV*, 2005.

[32] S. Yang, L. Bo, J. Wang, and L. G. Shapiro. Unsupervised template learning for fine-grained object recognition. In *NIPS*, 2012.

[33] B. Yao, G. Bradski, and L. Fei-Fei. A codebook-free and annotation-free approach for fine-grained image categorization. In *CVPR*, 2012.

[34] B. Yao, A. Khosla, and F.-F. Li. Combining randomization and discrimination for fine-grained image categorization. In *CVPR*, 2011.

[35] N. Zhang, R. Farrell, and T. Darrell. Pose pooling kernels for sub-category recognition. In *CVPR*, 2012.