Fine-Grained Recognition without Part Annotations

Jonathan Krause[1], Hailin Jin[2], Jianchao Yang[2], Li Fei-Fei[1]
[1]Computer Science Department, Stanford University. [2]Adobe Research.

The field of fine-grained recognition has made great progress in recognizing an ever-increasing number of categories. Compared to generic object recognition, fine-grained recognition benefits more from learning critical parts of objects that can help align objects of the same class and discriminate between neighboring classes. Current state-of-the-art results are, therefore, from models that require part annotations as part of a supervised training process [2, 9]. This poses a problem for scaling up fine-grained recognition to an increasing number of domains. In our work, we aim to alleviate this problem by developing a method for fine-grained recognition which forgoes the use of part annotations.

Our approach generates parts for recognition in an unsupervised fashion (Fig. 1). It begins by segmenting all images in the training set via co-segmentation. After finding a graph used to determine which images to align, we use the segmentations to align all images. Points sampled and propagated to all image form the basis for parts used in recognition.

Our co-segmentation formulation is similar to the "image+class" model of [4], with the addition of a unary foreground prior:

$$E(x_p^i; p_f) = \begin{cases} \log(p_f) & x_p^i = 1 \\ \log(1-p_f) & x_p^i = 0 \end{cases}, \quad (1)$$

where x_p^i is the binary segmentation assignment for pixel p in image i, and p_f is the strength of the prior. We use this in a refinement step of segmentation, where we do a binary search on p_f in order to find a segmentation spanning at least $\rho = 50\%$ the bounding box width and height, and between $\omega_1 = 10\%$ and $\omega_2 = 90\%$ the bounding box area. This foreground prior is useful for fixing a common failure mode of segmentation methods of either greatly over- or under-segmenting images, leading to an average improvement in Jaccard similarity of 7.69 on the CUB-2011 dataset [8].

We decompose the task of aligning objects in all images as the composition of aligning objects with similar poses, represented as a graph computed from disjoint minimum spanning trees on top of conv_4 features from a convolutional neural network, which work well to represent pose. Alignments between objects with similar poses are then done via shape context [1] on top of our segmentations. To generate parts, we sample points in the estimated foreground of a single image, then propagate them to all other images using the graph, forming a part out of a region around each point. At test time, we experiment with two methods of part detection. The first, a part detector method based on the Δ_{box} approach of Zhang *et al.* [9], trains an R-CNN [3] for the entire object and every part by treating each as a separate category. The second method is an extension of our segmentation and alignment procedure. After detecting the whole object with an R-CNN, we use the predicted bounding box in our segmentation framework and align the resulting segmentation with a set of nearest neighbors from the training set. Then, we propagate the parts from the training images to the test image.

We also present a method for discriminatively combining the information at each part together, which yields a significant improvement over simply concatenating the information at each part. Let f_p^i be features for image i at part p and $u_{c^i,c}^i(p)$ be the difference in decision values between incorrect class c^i and correct class c, determined by classifiers trained independently on each part p. Our goal is to learn a vector of weights v satisfying:

$$\min_v \sum_i \sum_{c \neq c^i} \max\left(0, 1 - v^T u_{c^i,c}^i\right)^2 + \lambda \|v\|_1 \quad (2)$$

which is equivalent to a one-class SVM (an SVM with only positive labels) with an L_2 loss and L_1 regularization. The final classification is determined by the combination of class decision values at each part, weighted by v.

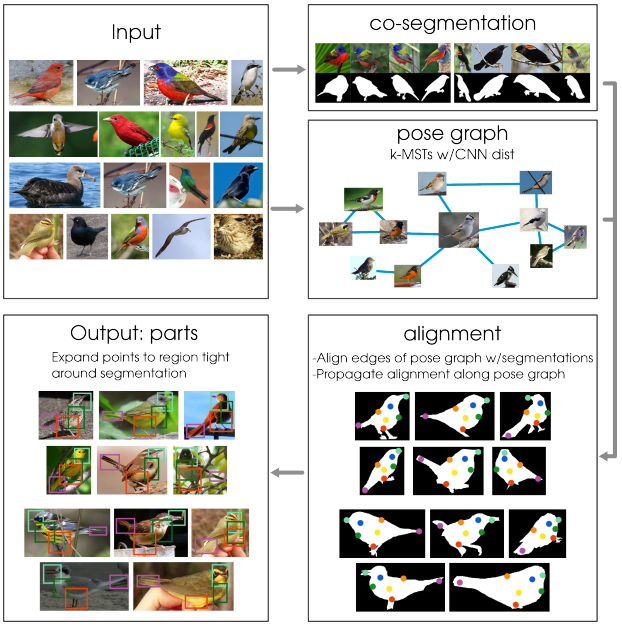

Figure 1: Overview of our method to generate parts used for recognition.

We do an extensive experimental evaluation of many variants of our method, including an evaluation of the CNN used for feature extraction (either a CaffeNet [5] or a VGGNet [7]). With a basic CaffeNet, our method is within 2% of state-of-the-art [2] on CUB-2011, despite using no part annotations, and with a stronger VGGNet our method is significantly better than any prior work (barring work that uses ground truth part locations *at test time*), bringing performance on CUB-2011 up to 82.0%. Since our method does not require part annotations, we can also evaluate on other datasets without the high level of annotation of CUB-2011. To this end, we evaluate on the cars-196 dataset [6] and set a new state-of-the-art by a large margin, bringing 196-way classification accuracy up to 92.6%.

[1] Serge Belongie, Jitendra Malik, and Jan Puzicha. Shape matching and object recognition using shape contexts. *PAMI*, 2002.

[2] Steve Branson, Grant Van Horn, Pietro Perona, and Serge Belongie. Improved bird species recognition using pose normalized deep convolutional nets. In *British Machine Vision Conference*, 2014.

[3] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[4] Matthieu Guillaumin, Daniel Küttel, and Vittorio Ferrari. Imagenet auto-annotation with segmentation propagation. *IJCV*, 2014.

[5] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[6] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *3dRR*, 2013.

[7] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[8] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD birds-200-2011 dataset. 2011.

[9] Ning Zhang, Jeff Donahue, Ross Girshick, and Trevor Darrell. Part-based R-CNNs for fine-grained category detection. In *ECCV*. 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.