Deformable Part Descriptors for Fine-grained Recognition and Attribute Prediction

Ning Zhang[1] Ryan Farrell[1,2] Forrest Iandola[1] Trevor Darrell[1]
[1]ICSI / UC Berkeley [2]Brigham Young University
[1]{nzhang,forresti,trevor}@eecs.berkeley.edu [2]farrell@cs.byu.edu

Abstract

Recognizing objects in fine-grained domains can be extremely challenging due to the subtle differences between subcategories. Discriminative markings are often highly localized, leading traditional object recognition approaches to struggle with the large pose variation often present in these domains. Pose-normalization seeks to align training exemplars, either piecewise by part or globally for the whole object, effectively factoring out differences in pose and in viewing angle. Prior approaches relied on computationally-expensive filter ensembles for part localization and required extensive supervision. This paper proposes two pose-normalized descriptors based on computationally-efficient deformable part models. The first leverages the semantics inherent in strongly-supervised DPM parts. The second exploits weak semantic annotations to learn cross-component correspondences, computing pose-normalized descriptors from the latent parts of a weakly-supervised DPM. These representations enable pooling across pose and viewpoint, in turn facilitating tasks such as fine-grained recognition and attribute prediction. Experiments conducted on the Caltech-UCSD Birds 200 dataset and Berkeley Human Attribute dataset demonstrate significant improvements over state-of-art algorithms.

1. Introduction

Despite the many important applications and domains under investigation, fine-grained recognition remains very challenging. As described in [23], what often differentiates *basic-level* categories is the presence or absence of parts (*e.g*. an elephant has 4 legs and a trunk), whereas *subordinate* categories are more often discriminated by subtle variations in the shape, size and/or appearance properties of these parts (*e.g*. elephant species can be distinguished by localized cues such as ear shape and size). Localizing and describing the object's parts therefore becomes central to uncovering its fine-grained identity.

Several approaches have been proposed for localizing and describing object parts in fine-grained domains. Pose-normalization was proposed for fine-grained recognition by Farrell *et al*. [23] and extended in Zhang *et al*. [47]. This paradigm seeks to discount variations in pose, articulation and camera viewing angle by localizing semantic object parts and extracting appearance features with respect to those localized parts. In these approaches part localization was accomplished using Poselets [11, 10], which require computationally-expensive filter ensembles for part localization as well as extensive supervision.

Parkhi *et al*. [36, 37] provide an alternate way of detecting basic-level category objects such as dogs and cats by using a Deformable Part Model [24] trained specifically on the head. Once a head has been detected in a test image, this region is used to initialize a grab-cut [38] segmentation and obtain a mask/silhouette of the entire object. Classification is then performed using a combination of the head shape and features extracted from within the head/body mask. While this method is relatively effective for front-facing cats and dogs, subjects in other domains (such as birds and people) often exhibit greater variation in pose and appearance and can be far more difficult to segment from their surroundings (see examples in Figure 1).

In this work, we introduce deformable part descriptors (DPD), a robust and efficient framework for pose-normalized description based on DPM parts and demonstrate its effectiveness for both fine-grained recognition and attribute prediction (see Figure 2). We propose two deformable part descriptors: the DPD-strong leverages the semantics inherent in strongly-supervised DPM parts; the DPD-weak exploits semantic annotations to learn cross-component correspondences, computing pose-normalized descriptors from weakly-supervised DPM parts. We present state-of-the-art results in evaluating our approach on standard fine-grained and domain-specific attribute datasets including the Caltech-UCSD birds dataset [13] and Human Attributes dataset [10]. An end-to-end open-source implementation of our method has been released with this paper and is available at http://dpd.berkeleyvision.org.

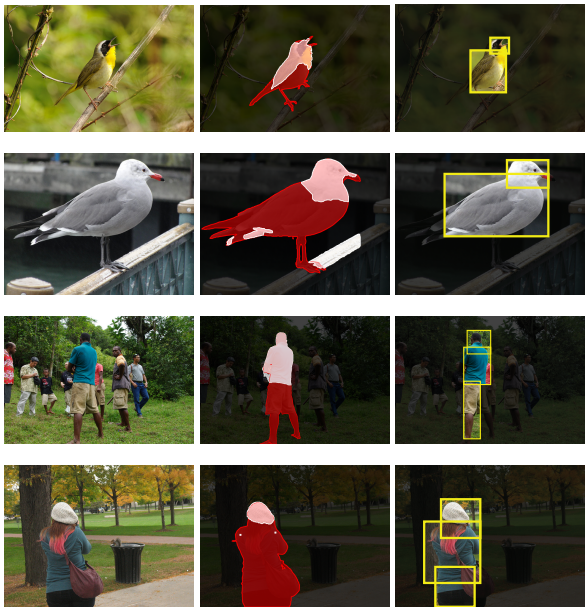

Figure 1. **Robustness of Deformable Part Descriptors (DPDs).** While they work well for generally homogeneous objects such as dogs and cats, head-seeded segmentation-based descriptors often fail for other domains such as birds and humans. Examples of this are shown in the middle column. The ground truth segmentation is overlaid in red; the grab-cut based segmentation [38], seeded with the head as foreground and everything outside the bounding box as background, is overlaid in white. The right column shows our part localization results using the strongly-supervised DPM. Best viewed in color.

2. Background

2.1. Fine-grained categorization

Fine-grained classification has recently emerged as a topic of great interest. A growing body of literature has proposed various techniques and has addressed recognition on a large number of fine-grained domains. These domains include: dog breed classification [26, 29, 37], subordinate categories of flowers [2, 1, 33, 34] and plants [4, 39], recognition of invertebrates [32], fine-grained classification on subsets of ImageNet [16] such as fungi [15], and species-level categorization of birds [23, 46]. We anticipate more work in the near future on man-made categories such as cars [40] and airplanes [30].

Following recent work by Belhumeur *et al*. [5] on localizing fiducial points in faces, Liu *et al*. [29] present an alternate approach to build an exemplar-based geometric and appearance models for detecting dog faces and localizing the facial keypoints. Another work in this vein is that of Sfar *et al*. [39] who classify leaves and flowers by describing appearance with multiple coordinate or "vantage" frames. A very recent and closely related contribution is the Part-based One-vs-One Features (POOF) proposed by Berg and Belhumeur [6] which leverages robust keypoint prediction learning a descriptor coordinate frame for each pair of keypoints.

Approaches such as [14, 36, 37] use region-level cues to estimate object segmentations which facilitate fine-grained classification. Some techniques use humans-in-the-loop, asking human annotators to click on object parts, answer questions regarding object attributes [13], or mark the region that best differentiates two confusing categories [17]. Template-based methods have also been investigated by Yao *et al*. [45] and by Yang *et al*. [43]. Such approaches use a set of fixed-position templates, thereby mitigating much of the computational cost of sliding window approaches, yet lack the spatial flexibility to deal with substantial pose variation.

2.2. Attribute Prediction

Attribute-based representations [22, 27, 28] present another promising direction, offering the possibility of recognizing a novel category with only a category description (no imagery). Relevant subsequent work on attributes includes that of Parikh and Grauman [35] exploring relative attribute strength and that of Berg *et al*. [7] and Duan *et al*. [19] propose automatic discover of attributes to aid in fine-grained classification. Perhaps the most closely related work on attribute prediction is Bourdev *et al*. [10], which uses features extracted from Poselet [11, 9] activations to predict nine binary attributes on images of humans. Another promising alternative to Poselet-based approaches for effective transfer of pose annotation from training images to test images is the Exemplar SVM proposed by Malisiewicz *et al*. [31].

2.3. Deformable Parts Model

Inspired by the Pictorial Structures work of Fischler and Elschlager [25], the Deformable Parts Model (DPM) of Felzenszwalb *et al*. [24] has become one of the most effective and widely-used object detection approaches to date. The object is represented by a coarse root HOG filter and several higher resolution part filters. The DPM model uses a mixture of components to capture variation in viewpoint and/or pose (*e.g*. for a car, the three components might correspond roughly to the front, side and three-quarter views).

Strongly-supervised DPM While the limited supervision required for the DPM is advantageous, the latent parts provide no semantic information about the object which makes pose-normalization challenging. Related work has used strong supervision to train DPMs for human pose estimation [41, 44], part localization [12] and object detection [3]. While these methods focus on detection, our objective is different, pooling semantic part features across components to derive a pose-normalized representation.

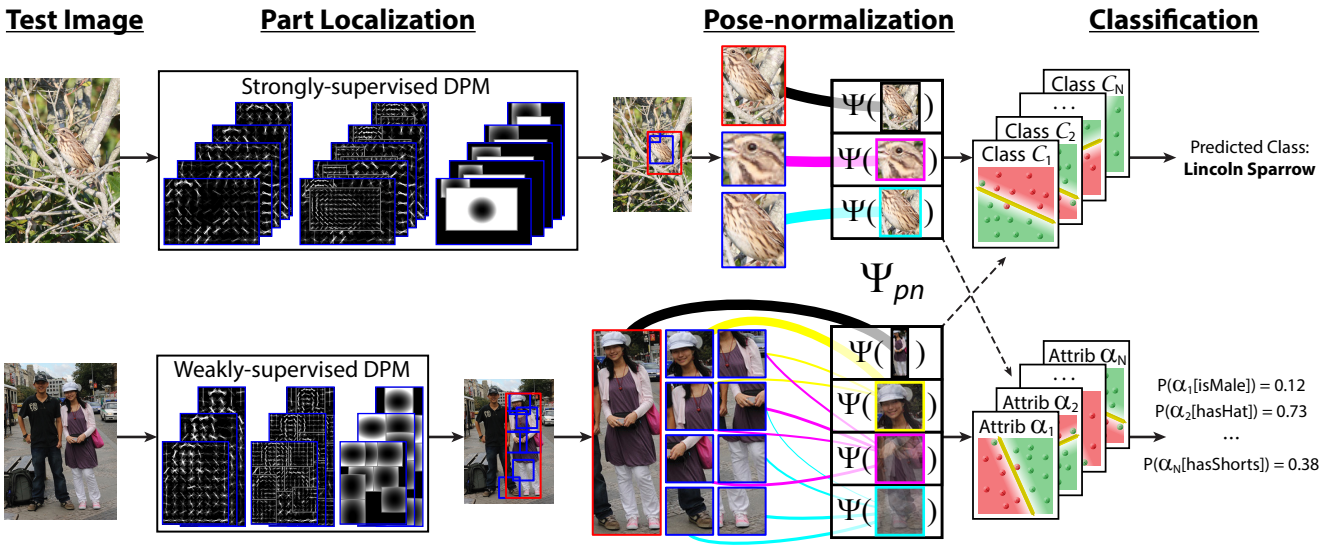

Figure 2. **Overview.** Both of the proposed Deformable Part Descriptors (DPDs) are depicted above. The first descriptor (top row) applies a strongly-supervised DPM [3] for part localization and then performs pose-normalization by pooling features from these inherently semantic parts. The pose-normalized features Ψ_{pn} in Equation 3 can be used for either fine-grained recognition or attribute prediction (as indicated by the dotted arrows). The second descriptor employs a weakly-supervised DPM [24] for part localization and then uses a learned semantic correspondence model to pool features from the component-specific localized latent parts into semantic regions (the pose-normalized descriptor Ψ_{pn}) shared by all components. In the part to region pooling (the black, yellow, magenta, and cyan lines), wider lines indicate higher weight. Best viewed in color.

In this work, we adopt a variant of the strongly-supervised DPM [3]. We use object part annotations and aim to localize the semantic parts for pose-normalized descriptors. Instead of initializing the fixed-size part filters heuristically (energy maximization), we initialize the part filters by using semantic part annotations. As in [3], we initialize the mixture components by clustering the annotated poses instead of using the aspect ratio. After training the mixture of root filters, we process the training images to get the component assignment for each training example. Then, for each component c, the part filters of this component are initialized at the average relative locations with average size among all the examples with component assignment c. Unlike [3], we do not impose constraints on part overlap during training.

Notation For both types of DPM, we use a latent SVM to train the model parameters. The matching score of a model β for a given image $\mathbf{I}$ is

$$S(I,\beta) = F_0^\beta \cdot \Phi(I,p_0) + \sum_{i=1}^{n} S_{\text{part}}(p_i,I,\beta) + b$$
$$S_{\text{part}}(p_i,I,\beta) = F_i^\beta \cdot \Phi(I,p_i) - d_i \cdot \Phi_d(dx_i,dy_i) \quad (1)$$

where F_0^β is the root filter, F_i^β is the part filter, b is the bias term, $\Phi(I,p)$ is the image feature vector of the sub-window located at position p, (dx_i,dy_i) and d_i are respectively is the displacement and displacement weight of the i-th part, and $\Phi_d(dx,dy) = (dx,dy,dx^2,dy^2)$ is used to penalize the part displacement[24]. The overall score is computed using the best possible placements, i.e. $f_\beta = max_{z\in Z(I)} S(I,\beta)$ where $Z(I)$ is the set of all possible latent values. The model is trained by iteratively assigning the best possible placements and learning the parameter β by stochastic gradient descent. Then, the part locations are predicted by

$$Z(I) = argmax_{(Z,\beta)} S(I,\beta) \quad s.t.\ O(p_0(I),\text{bbox}) > \delta$$
$$Z(I) = (c(I),p_0(I),\cdots,p_n(I)) \quad (2)$$

where $c(I)$ is the component assignment, $p_0(I)$ is the predicted bounding box, $p_i(I)$ is the predicted location of part p_i and $O(p_0,bbox)$ is the overlap between the predicted bounding box and ground truth bounding box. δ is a threshold to control predictions and we set it to be 0.65 in our experiments. Given the part localizations, we show below how to form DPM-based pose-normalized descriptors and pool them across components, in a manner analogous to what the Pose-pooling Kernel (PPK) [47] does via Poselets.

3. Deformable Part Descriptors (DPD)

We use the deformable part model (DPM)'s demonstrated effectiveness for detecting objects as a foundation for our pose-normalized approach to fine-grained recognition and attribute prediction. We do this as a faster and more reliable alternative to Poselets which were previously proposed for pose-normalization in both fine-grained recognition [23, 47] and attribute prediction [10]. As noted earlier,

we are not the first to consider using DPMs for fine-grained recognition. Parkhi *et al*. [36, 37] trained a DPM model to locate a cat's or a dog's head and then used this detection both to describe the head appearance and to seed a segmentation which would recover the rest of the body (See Figure 1). Our goal is to use DPM to localize the parts and pool the pose-normalized image features induced by the part locations.

3.1. Strongly-supervised DPD

The underlying principle in pose-normalization is that one can decompose an object's appearance as observed in one image and compare it to the same object (or object category) as observed in a different image. This decomposition generally equates to localizing semantic parts and then describing them.

From the DPM models, suppose that we have a total of C components including mirror pairs $\mathcal{C} = \{c^{(1)}, c^{(2)}, \ldots, c^{(C)}\}$, where for odd values of j, $c^{(j+1)}$ is the mirror component of $c^{(j)}$. Each component $c^{(j)}$ has a set of parts $\mathcal{P}^{(j)} = \{p_1^{(j)}, p_2^{(j)}, \ldots, p_P^{(j)}\}$, $p_i^{(j)}$ denoting the ith part for component $c^{(j)}$. We now define a pose-normalized representation with R semantic pooling regions $\mathcal{R} = \{r_1, r_2, \ldots, r_R\}$. We define the pose-normalized representation as

$$\Psi_{pn}(I) = \left[\Psi\left(I, r_0\right), \Psi\left(I, r_1\right), \ldots, \Psi\left(I, r_R\right)\right]. \quad (3)$$

where $\Psi(I, r_l)$ is the pooled image feature for semantic region r_l and $\Psi(I, r_0)$ is the image feature inside the root filter or bounding box and we concatenate the image features together to get the final pose-normalized representation Ψ_{pn}.

To derive the pose-normalized representation for a given detection $Z(I)$ in Equation 2 (from component $c^{(j)} = c(I)$), we must figure out a mapping $\mathcal{S}^{(j)} : \mathcal{P}^{(j)} \rightarrow \mathcal{R}$. For each part $p_i^{(j)}$, we seek to determine which pose-normalized pooling region or regions $\{r_l\}$ the features $\Psi(I, p_i^{(j)})$ should be mapped into. For strongly-supervised DPM, we use the semantic part annotations and it is straightforward to pool the corresponding part descriptor across different components by setting the semantic regions the same as the semantic parts from the strong DPM, i.e. $\mathcal{R} = \mathcal{P}^{(j)}$ for all $j \in \{1, \ldots, C\}$.

$$\Psi(I, r_l) = \Psi(I, p_l^{(j)}) \qquad j \in \{1, \ldots, C\} \quad (4)$$

3.2. Weakly-supervised DPD

Using a weak DPM, the latent parts of different components have no explicit semantic correlation, nor is such correspondence guaranteed to exist. We have explored two options for dealing with this: one based on a per-component appearance representation with no semantic parts; the other leverages additional annotations to estimate semantic correspondence of the latent parts across components.

The per-component representation is very straightforward. The training and test sets are partitioned into subsets according to which component fires on them. A separate classification model is trained for each component using the features extracted on that component's subset of training images. This creates reasonable classifiers, however, there are two shortcomings. First, the parts carry no semantic information (though this is not inherently necessary). Second, the training data is fragmented across the C components, so models will be accordingly weaker. Training fragmentation is particularly problematic for fine-grained recognition where you may only have 15-30 training examples total for each category. Experiments using this model are included in Section 4.3.

The second representation, however, solves both of these problems. By providing semantic annotations at training time, we can model semantic correspondence between the latent parts of different components. In effect, we learn the semantic identity of each latent part in the model and can thus pool features to a global *pose-normalized* model, independent of which component fires during detection.

We model the pose-normalization as a weighted bipartite graph $G = (\mathcal{P}, \mathcal{R}, \mathcal{W})$ where $w_{il}^{(j)} \in \mathcal{W}$ indicates the degree to which $p_i^{(j)}$, the i-th part of component $c^{(j)}$ contributes to r_l.

$$\Psi\left(I, r_l\right) = \frac{1}{N} \sum_{i=1}^{P} w_{il}^{(j)} \cdot \Psi\left(I, p_i^{(j)}\right) \quad (5)$$

where $N = \sum_i^P w_{il}^{(j)}$ is used for normalization. $\mathcal{W}$ is a three dimensional matrix with size $|\mathcal{P}| \times |\mathcal{R}| \times |\mathcal{C}|$ and $w_{il}^{(j)}$ is modeled as a function of the annotations $\mathcal{A}$. Annotations $a_k \in \mathcal{A}$ could be keypoints (as used to train Poselets), rectangular regions (as used to train Strong DPM) or any other type of semantic labels. In our example, we use the keypoint annotations included with the CUB2011 dataset [42] and H3D dataset [11]. More precisely, we define $w_{il}^{(j)}$ as:

$$w_{il}^{(j)} = \sum_{k=1}^{A} \rho_{kl} \cdot overlap\left(a_k, p_i^{(j)}\right) \quad (6)$$

where $\rho_{kl} \in [0, 1]$ indicates a specified semantic relevance for annotation a_k to region r_l (*e.g.* left ear is relevant to head, left knee is not). The $overlap\left(a_k, p_i^{(j)}\right)$ function encodes the distribution of annotation a_k within part $p_i^{(j)}$.

Let $\mathcal{I}_{jk}$ be the set of training images that have semantic annotation a_k and where $c^{(j)}$ is the component which fires. We can formally define the fractional overlap

$$overlap\left(a_k, p_i^{(j)}\right) = \frac{|\{I \in \mathcal{I}_{jk} | a_k(I) \cap p_i^{(j)} \neq \emptyset\}|}{|\mathcal{I}_{jk}|}. \quad (7)$$

It is worth noting that the training set need not be exhaustively labeled with semantic annotations. Fewer annotations

just mean smaller $\mathcal{I}_{jk}$ and thus coarser predictions of the weights $w_{il}^{(j)}$.

3.3. Classification

Given the pose-normalized representation $\Psi_{pn}(I)$ in Equation 3, we employ a linear SVM for the final classification. For the strongly-supervised DPD, we use the annotated semantic parts to get Ψ_{pn} for training and part localization from Equation 2 for testing; for the weakly-supervised DPD, we utilize Ψ_{pn} from Equation 2 for both training and testing. Due to the sparsity of training examples for certain poses, there will be some test images for which a correct detection cannot be found even given a specified object bounding box. This means that we don't have predictions for the part locations in Equation 2. In such cases, we can use the classifier trained on the feature descriptor inside the bounding box $\Psi(I, r_0)$ without any pose-normalization.

4. Experiments

In this section, we will present a comparative performance evaluation of our proposed method. We conduct experiments on the commonly used fine-grained benchmark Caltech-UCSD bird dataset [42] as well as the Berkeley Human Attribute dataset from [10]. Our system can process several images per second, leveraging the available fast DPM implementation in [20]. An open-source version of our end-to-end DPD implementation is available at `http://dpd.berkeleyvision.org`. In contrast, the fastest available implementation of Poselet-based [47] relies on a C++ reimplementation of [11] and takes approximately 10 seconds per frame on a comparable machine.

4.1. Implementation Details

Image Features and Classification After predicting the part regions via DPM, we use kernel descriptors [8] to extract feature vectors for final classification.[1] Specifically, we use four types of kernel descriptors: gradient, local binary pattern (lbp), rgb color, and normalized rgb color. Following the setting in [8], we compute kernel descriptors on local image patches of size 16 x 16 over a dense regular grid of step size 8 and then apply a spatial pyramid on top. We use vector quantization of these descriptors with a 1000-element codebook, concatenating per-region descriptor histograms into a single vector per image. These vectors are provided as input to a linear support vector machine (`liblinear` with power scaled features).

Efficient Weak Pooling In Equation 6, while a fully optimal formulation would likely learn the best convex combination of ρ_{kl} and utilize sophisticated distributions for all pairs $(a_k, p_i^{(j)})$, we make two simplifying assumptions for efficiency and simplicity. First, we define ρ_{kl} as an indicator such that $\rho_{kl} \in \{0, 0.5, 1\}$ and $\sum_l \rho_{kl} = 1$; in other words, each semantic part a_k is associated with one or more regions r_l (in a few cases it is shared between two, such as the shoulder between head and torso regions). This represents partitioning the semantic parts $\mathcal{A}$ amongst the pooling regions $\mathcal{R}$. Second, to evaluate the $overlap(a_k, p_i^{(j)})$ function in Equation 7, we evaluate the weakly-supervised detector on the semantically annotated data, noting for each detection where the various parts $\{a_k\}$ fall. Across the annotated data we accumulate distributions for each part $p_i^{(j)}$ (we only include contributions toward $p_i^{(j)}$ when $c^{(j)}$ fires). While the optimal method would likely model these spatial distributions with respect to $p_i^{(j)}$, we relax this and simply record the fraction of semantically annotated images which have a_k, for which $c^{(j)}$ fires and where a_k has high overlap with (or falls within) $p_i^{(j)}$.

[1] For results using deep convolutional features, see [18] for more details.

4.2. Caltech-UCSD Birds-200 Dataset

We conduct our experiments on the 200-category Caltech UCSD bird dataset, one of the most widely used and competitive fine-grained classification benchmarks. Following [23], we use two semantic parts for the bird dataset: head and body. We utilize both the earlier (CUB200-2010) and current (CUB200-2011) versions of the bird dataset for better comparison.

CUB200-2010 The initial version of the CUB200 dataset contains 6033 images of 200 different bird species in North America. There are around 30 images per class. We use the provided train-test split which uses 15 images per class for training and the balance for testing. Each image in the dataset has a bounding box annotation but no other annotation is provided. In order to train the strong DPM with semantic parts, we manually annotate the part locations for the training images, i.e. head and body boxes. If one of the parts is not visible, we mark its visibility state as 0. Given those annotations on training images, we train a strong DPM with 5 components, each with these two parts. Due to the lack of additional annotations, we are unable to evaluate the DPD-weak approach on this dataset.

Table 1 shows the mean accuracy of our method on the CUB200-2010 dataset. We also compare with other published methods, including MKL [13], random forest [46], TriCos [14], template matching [43], segmentation [1] and the recently-published Bubblebank method [17]. One baseline is to use the same image descriptors used by our methods inside the bounding box but without any pose normalization, i.e. KDES [8], which yields 26.4% mean accuracy, better than some previous methods. Our approach achieves 34.5% mean accuracy, which outperforms the best previous

Method	Mean Accuracy(%)
MKL [13]	19.0
Random Forest [46]	19.2
KDES [8]	26.4
TriCos [14]	26.7
Template matching [43]	28.2
Segmentation [1]	30.2
Bubblebank [17]	32.5
DPD-strong-2	**34.5**

Table 1. **Results on CUB200-2010 dataset.** We note that [17] uses additional human labels. The best performance is achieved by the two-part strongly-supervised DPD model (DPD-strong-2).

Method	Mean Accuracy(%)
KDES [8]	42.53
Template matching[43]	43.67
Oracle	64.53
DPD-weak-8	**50.98**
DPD-strong-2	50.05
DPD-strong-2-no-head	43.15
DPD-strong-2-no-body	46.77

Table 2. **Results on CUB200-2011 dataset using KDES features.** The best performance is achieved by the 8-part weakly-supervised DPD model (DPD-weak-8). Oracle is akin to DPD-strong but uses the ground truth head and body locations.

result [17], and that requires human annotations obtained using a crowdsourced online game to find the most discriminative regions.

CUB200-2011 The CUB200-2011 is the latest version of the dataset, which includes more high-quality images and has 15 part annotations, e.g. beak, crest, throat, left-eye, left-wing, nape, etc. This dataset contains 11,788 images of the same 200 bird species as CUB200-2010. We use the default training/test split, which gives us around 30 training examples per class. We train both a weak DPM and a strong DPM to facilitate part localization. For the strong DPM, we train a mixture of five components, each with two parts and we partition the keypoint annotations to generate the head and body part annotations. Specifically, we choose the semantic part annotations as the minimum rectangular region to cover the part's specified subset of keypoints. This strong DPM enables part localization with the following accuracies: the head part with 45.1% precision and 43.5% recall, the body part with 77.5% precision and 75.2% recall. Here we mark a correct prediction when the predicted part box and the ground truth part box have an overlap over 0.5. For the weak DPM, we train a mixture of three components, each with 8 parts and using pooling as described in Equation 5. The pooling weights learned for the bird model are

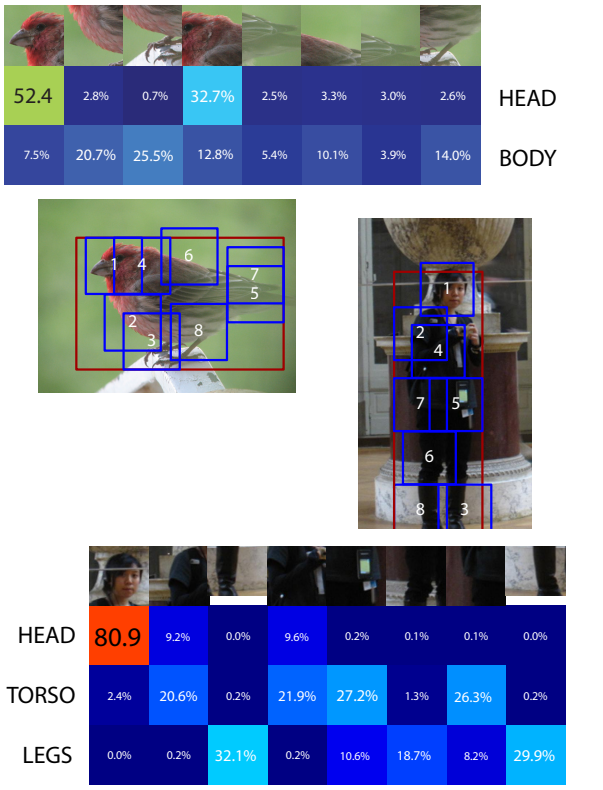

Figure 3. **Learned Per-component Pooling Weights.** Pictured are examples of the weighting models that are learned for the weakly-supervised bird (above) and human (below) models (DPD-weak). Shown is just one of the components for each model. In each weight matrix, rows correspond to the semantic regions that the eight latent parts (columns) are pooled to. We emphasize that the weights are learned automatically. Best viewed in color.

visualized in the top part of Figure 3.

Table 2 presents our classification results on the CUB200-2011 dataset using KDES [8] features. To compute the baseline method, we use only the bounding box region, yielding 42.53% mean accuracy. We also evaluate the template matching method of [43], using code released by the authors. Our DPD-strong achieves 50.05% mean accuracy and DPD-weak achieves 50.98% on this dataset, outperforming other state-of-art methods. To understand the contribution of individual DPM parts to the classification accuracy, we blind a DPD-strong to individual semantic parts in an ablation study. These results are included in Table 2. We find that localizing the head part of the bird is especially important: removing the head part, the CUB200-2011 accuracy falls to 43.15%. Additional experiments on the importance of individual parts are available on our DPD project webpage. In terms of other published work on this dataset, PPK [47] makes use of Poselet activations to generate pose-normalized representation and achieves a

Attribute	Freq	SPM [10]	Poselets [10]	Per-component	DPD-weak-8	DPD-strong-3
is male	0.593	68.1	82.4	80.5	82.9	**83.7**
has long hair	0.300	40.0	**72.5**	60.8	67.8	70.0
has glasses	0.220	25.9	**55.6**	33.6	40.7	38.1
has hat	0.166	35.3	60.1	61.3	70.3	**73.4**
has t-shirt	0.235	30.6	**51.2**	43.7	46.1	49.8
has long sleeves	0.490	58.0	74.2	74.3	76.5	**78.1**
has shorts	0.179	31.4	45.5	50.3	59.4	**64.1**
has jeans	0.338	39.5	54.7	72.3	77.1	**78.1**
has long pants	0.747	84.3	90.3	90.6	93.0	**93.5**
Mean AP	0.363	45.91	65.18	63.03	68.20	**69.88**

Table 3. **Results on the Human Attributes dataset.** Freq is the label frequency (fraction of specified attributes that are positive) and Per-component gives the results when using the per-component pooling method discussed in the beginning of Section 3.2. The values above denote mean average precision. For the full Precision-Recall curves, see Figure 4. Another baseline using the same image features but considering only the bounding box region achieves 66.58% mean AP.

mean accuracy of 28.18%. The authors of POOF [6] recently reported performance of 56.89% enabled both by accurate part (keypoint) localization and through the training of thousands of classifiers. We have experimented with replacing the KDES features with deep convolutional features and have reported 64.96% performance (please see [18] for details).

4.3. Human Attributes Dataset

The human attributes dataset [10] contains 8035 images collected from the H3D [11] and PASCAL VOC 2010 [21] datasets. There are nine attributes (*e.g.* has t-shirt, has jeans), and each one has a label of {-1, 1, 0} respectively meaning absent, present and unspecified. We use the H3D data to train a strong DPM with 3 components and 3 semantic parts (head, torso and legs) and a weak DPM with 3 components and 8 parts. Example pooling weights learned for the weak DPM are visualized in the lower part of Figure 3. Similar to the CUB200-2011 dataset, the semantic part annotations are generated from the keypoint annotations available from H3D. We then test on both the training and test images of the human attributes dataset to localize the parts and get pose-normalized image descriptors for attribute prediction. Table 3 shows prediction results on these nine attributes. Two baselines are included for comparison: SPM which uses spatial pyramid match inside the bounding box and the Poselet approach of Bourdev, *et al.*; both results are taken from [10]. We also include the results of using per-component classifiers, measuring average precision under the precision-recall curve. Precision-Recall curves for each of the nine attributes are shown in Figure 4. We use mean AP instead of mean accuracy for this dataset because the percentage of positive examples for each attribute is quite varied. From the table, we see that both DPD methods outperform pose-normalization using Poselets and moreover, our method is approximately 30 times more efficient.

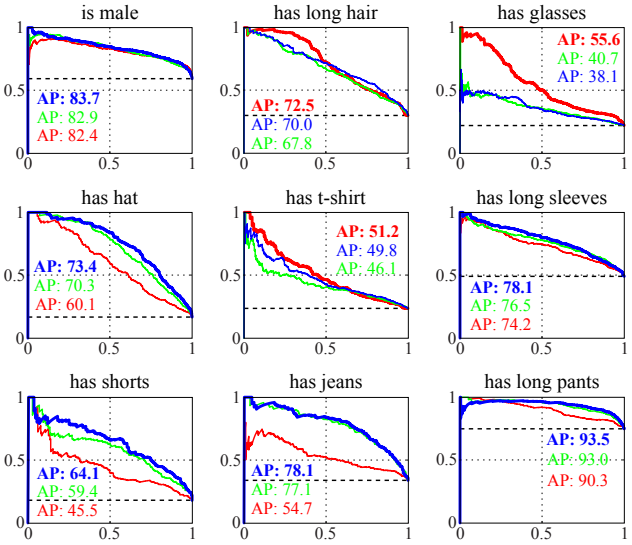

Figure 4. **Attribute Prediction Precision-Recall Curves.** Results are shown for each given attribute. The blue curve and mean average precision (AP) scores are for the DPD-strong. The green curve and mean AP scores are for the DPD-weak. The red curve and mean AP scores are for the previous start-of-the art technique of Bourdev *et al.* [10]. The dashed line indicates the frequency (fraction of positives). Best viewed in color.

5. Conclusion

In this paper we have proposed Deformable Part Descriptors (DPDs), a pose-normalized representation based on DPMs. We described two such pose-normalized methods, respectively applicable to strongly-supervised and weakly-supervised variants of deformable part models. The first method exploits the semantics inherent in the strongly-supervised DPM's parts, pooling them directly to form a pose-normalized descriptor. The second uses semantic an-

notations to learn cross-component correspondences between parts of the weakly-supervised DPM. These correspondences are then used to generate a pose-normalized descriptor. We have evaluated the proposed DPD methods, surpassing the previous state-of-the-art performance on standard datasets for both fine-grained recognition and attribute prediction.

In conclusion, we outline some directions for future work. First, we suggest that a greater number of supervised parts (as used in Azizpour *et al*. [3]) would increase the descriptive power of the DPD model. However, to do this, we would need to address the issue of self-occlusion. Second, learning of cross-component part correspondences could be enhanced by considering unconstrained convex combinations for the semantic relevance coefficients ρ_{kl} and more optimal $overlap(\cdot)$ functions that address the spatial distributions of the semantic annotations within parts, instead of simply considering occurrence frequency.

Acknowledgments This research is funded by NSF grants IIS-1116411 and IIS-1212928, by DARPA's Minds Eye and MSEE programs, and by the Toyota Motor Corporation. The third author is supported by the NDSEG Fellowship program.

References

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

[2] A. Angelova, S. Zhu, and Y. Lin. Image segmentation for large-scale subcategory flower recognition. In *WACV*, 2013.

[3] H. Azizpour and I. Laptev. Object Detection Using Strongly-Supervised Deformable Part Models. In *ECCV*, 2012.

[4] P. N. Belhumeur, D. Chen, S. Feiner, D. Jacobs, W. J. Kress, H. Ling, I. Lopez, R. Ramamoorthi, S. Sheorey, S. White, and L. Zhang. Searching the Worlds Herbaria: a System for Visual Identification of Plant Species. In *ECCV*, 2008.

[5] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing Parts of Faces Using a Consensus of Exemplars. In *CVPR*, 2011.

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

[7] T. Berg, A. Berg, and J. Shih. Automatic Attribute Discovery and Characterization from Noisy Web Data. In *ECCV*, 2010.

[8] L. Bo, X. Ren, and D. Fox. Kernel Descriptors for Visual Recognition. In *NIPS*, 2010.

[9] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting People Using Mutually Consistent Poselet Activations. In *ECCV*, 2010.

[10] L. Bourdev, S. Maji, and J. Malik. Describing People: Poselet-Based Approach to Attribute Classification. In *ICCV*, 2011.

[11] L. Bourdev and J. Malik. Poselets: Body Part Detectors Trained Using 3D Human Pose Annotations. In *ICCV*, 2009.

[12] S. Branson, P. Perona, and S. Belongie. Strong Supervision From Weak Annotation: Interactive Training of Deformable Part Models. In *ICCV*, 2011.

[13] 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.

[14] 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.

[15] J. Deng, A. Berg, K. Li, and L. Fei-Fei. What Does Classifying More Than 10,000 Image Categories Tell Us? In *ECCV*, 2010.

[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

[17] J. Deng, J. Krause, and L. Fei-Fei. Fine-Grained Crowdsourcing for Fine-Grained Recognition. In *CVPR*, 2013.

[18] 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 *arXiv*, 2013.

[19] K. Duan, D. Parkh, D. Crandall, and K. Grauman. Discovering Localized Attributes for Fine-grained Recognition. In *CVPR*, 2012.

[20] C. Dubout and F. Fleuret. Exact Acceleration of Linear Object Detectors. In *ECCV*, 2012.

[21] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *IJCV*, 88(2), June 2010.

[22] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing Objects by their Attributes. In *CVPR*, 2009.

[23] R. Farrell, O. Oza, N. Zhang, V. I. Morariu, T. Darrell, and L. S. Davis. Birdlets: Subordinate Categorization using Volumetric Primitives and Pose-normalized Appearance. In *ICCV*, 2011.

[24] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object Detection with Discriminatively Trained Part Based Models. *PAMI*, 32(3), 2010.

[25] M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computers*, 22(1):67–92, January 1973.

[26] A. Khosla, N. Jayadevaprakash, B. Yao, and L. Fei-Fei. Novel dataset for fine-grained image categorization. In *FGVC Workshop, CVPR*, 2011.

[27] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and Simile Classifiers for Face Verification. In *ICCV*, 2009.

[28] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to Detect Unseen Object Classes by Between-Class Attribute Transfer. In *CVPR*, 2009.

[29] J. Liu, A. Kanazawa, D. Jacobs, and P. Belhumeur. Dog Breed Classification Using Part Localization. In *ECCV*, 2012.

[30] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-Grained Visual Classification of Aircraft. Technical report, 2013.

[31] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of Exemplar-SVMs for Object Detection and Beyond. In *ICCV*, 2011.

[32] G. Martinez-Munoz, N. Larios, E. Mortensen, W. Zhang, A. Yamamuro, R. Paasch, N. Payet, D. Lytle, L. Shapiro, S. Todorovic, A. Moldenke, and T. Dietterich. Dictionary-free categorization of very similar objects via stacked evidence trees. In *CVPR*, 2009.

[33] M.-E. Nilsback and A. Zisserman. A Visual Vocabulary for Flower Classification. In *CVPR*, 2006.

[34] M.-E. Nilsback and A. Zisserman. Automated Flower Classification over a Large Number of Classes. In *ICVGIP*, 2008.

[35] D. Parikh and K. Grauman. Relative Attributes. In *ICCV*, 2011.

[36] O. M. Parkhi, A. Vedaldi, C. V. Jawahar, and A. Zisserman. The Truth About Cats and Dogs. In *ICCV*, 2011.

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

[38] C. Rother, V. Kolmogorov, and A. Blake. GrabCut -Interactive Foreground Extraction using Iterated Graph Cuts. In *SIGGRAPH*, 2004.

[39] A. R. Sfar, N. Boujemaa, and D. Geman. Vantage Feature Frames For Fine-Grained Categorization. In *CVPR*, 2013.

[40] M. Stark, J. Krause, B. Pepik, D. Meger, J. J. Little, B. Schiele, and D. Koller. Fine-Grained Categorization for 3D Scene Understanding. In *BMVC*, 2012.

[41] M. Sun and S. Savarese. Articulated Part-based Model for Joint Object Detection and Pose Estimation. In *ICCV*, 2011.

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

[43] S. Yang, L. Bo, J. Wang, and L. Shapiro. Unsupervised Template Learning for Fine-Grained Object Recognition. In *NIPS*, 2012.

[44] Y. Yang and D. Ramanan. Articulated Pose Estimation using Flexible Mixtures of Parts. In *CVPR*, 2011.

[45] B. Yao, G. Bradski, and L. Fei-Fei. A Codebook-Free and Annotation-Free Approach for Fine-grained Image Categorization. In *CVPR*, 2012.

[46] B. Yao, A. Khosla, and L. Fei-Fei. Combining Randomization and Discrimination for Fine-grained Image Categorization. In *CVPR*, 2011.

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