

## Actions and Attributes from Wholes and Parts

Georgia Gkioxari  
UC Berkeley

gkioxari@berkeley.edu

Ross Girshick  
Microsoft Research

rbg@microsoft.com

Jitendra Malik  
UC Berkeley

malik@berkeley.edu

### Abstract

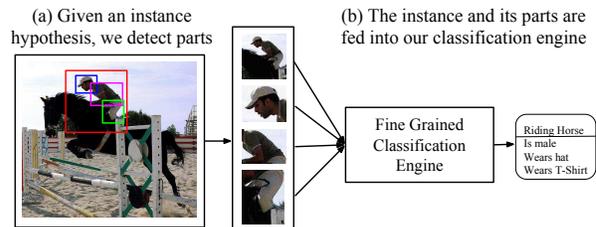
We investigate the importance of parts for the tasks of action and attribute classification. We develop a part-based approach by leveraging convolutional network features inspired by recent advances in computer vision. Our part detectors are a deep version of poselets and capture parts of the human body under a distinct set of poses. For the tasks of action and attribute classification, we train holistic convolutional neural networks and show that adding parts leads to top-performing results for both tasks. We observe that for deeper networks parts are less significant. In addition, we demonstrate the effectiveness of our approach when we replace an oracle person detector, as is the default in the current evaluation protocol for both tasks, with a state-of-the-art person detection system.

### 1. Introduction

For the tasks of human attribute and action classification, it is difficult to infer from the recent literature if part-based modeling is essential or, to the contrary, obsolete. Consider action classification. Here, the method from Oquab *et al.* [21] uses a holistic CNN classifier that outperforms part-based approaches [20, 30]. Turning to attribute classification, Zhang *et al.*'s CNN-based PANDA system [33] shows that parts bring dramatic improvements over a holistic CNN model. How should we interpret these results? We aim to bring clarity by presenting a single approach for both tasks that shows consistent results.

We develop a part-based system, leveraging convolutional network features, and apply it to attribute and action classification. For both tasks, we find that a properly trained holistic model matches current approaches, while parts contribute further. Using deep CNNs we establish new top-performing results on the standard PASCAL human attribute and action classification benchmarks.

Figure 1 gives an outline of our approach. We compute CNN features on a set of bounding boxes associated with the instance to classify. One of these bounding boxes corresponds to the whole instance and is either provided by an



**Figure 1:** Schematic overview of our overall approach. (a) Given an R-CNN person detection (red box), we detect parts using a novel, deep version of poselets (Section 3). (b) The detected whole-person and part bounding boxes are input into a fine-grained classification engine to produce predictions for actions and attributes (Section 4).

oracle or comes from a person detector. The other bounding boxes (three in our implementation) come from poselet-like part detectors.

Our part detectors are a novel “deep” version of poselets. We define three human body parts (head, torso, and legs) and cluster the keypoints of each part into several distinct poselets. Traditional poselets [1, 3] would then operate as sliding-window detectors on top of low-level gradient orientation features, such as HOG [7]. Instead, we train a sliding-window detector for each poselet on top of a deep feature pyramid, using the implementation of [13]. Unlike HOG-based poselets, our parts are capable of firing on difficult to detect structures, such as sitting versus standing legs. Also, unlike recent deep parts based on bottom-up regions [31], our sliding-window parts can span useful, but inhomogeneous regions, that are unlikely to group together through a bottom-up process (*e.g.*, bare arms and a t-shirt).

Another important aspect of our approach is task-specific CNN fine-tuning. We show that a fine-tuned holistic model (*i.e.*, no parts) is capable of matching the attribute classification performance of the part-based PANDA system [33]. Then, when we add parts our system outperforms PANDA. This result indicates that PANDA’s dramatic improvement from parts comes primarily from the weak holistic classifier baseline used in their work, rather than from the parts themselves. While we also observe an improvement from

adding parts, our marginal gain over the holistic model is smaller, and the gain becomes even smaller as our network becomes deeper. This observation suggests a possible trend: as more powerful convolutional network architectures are engineered, the marginal gain from explicit parts may vanish.

As a final contribution, we show that our system can operate “without training wheels.” In the standard evaluation protocol for benchmarking attributes and actions [2, 8], an oracle provides a perfect bounding box for each test instance. While this was a reasonable “cheat” a couple of years ago, it is worth revisiting. Due to recent substantial advances in detection performance, we believe it is time to drop the oracle bounding box at test time. We show, for the first time, experiments doing just this; we replace ground-truth bounding boxes with person detections from a state-of-the-art R-CNN person detector [12]. Doing so only results in a modest drop in performance compared to the traditional oracle setting.

## 2. Related Work

**Low-level image features.** Part-based approaches using low-level features have been successful for a variety of computer vision tasks. DPMs [9] capture different aspects of an object using mixture components and deformable parts, leading to good performance on object detection and attribute classification [32]. Similarly, poselets [1, 3, 14, 15, 20, 28] are an ensemble of models that capture parts of an object under different viewpoints and have been used for object detection, action and attribute classification and pose estimation. Pictorial structures and its variants [10, 11, 29] explicitly model parts of objects and their geometric relationship in order to accurately predict their location.

**Convolutional network features.** Turning away from hand-designed feature representations, convolutional networks (CNNs) have shown remarkable results on computer vision tasks, such as digit recognition [19] and more recently image classification [18, 23]. Girshick *et al.* [12] show that a holistic CNN-based approach performs significantly better than previous methods on object detection. They classify region proposals using a CNN fine-tuned on object boxes. Even though their design has no explicit part or component structure, it is able to detect objects under a wide variety of appearance and occlusion patterns.

**Hybrid feature approaches.** Even more recently, a number of methods incorporate HOG-based parts into deep models, showing significant improvements. Zhang *et al.* [33] use HOG-poselet activations and train CNNs, one for each poselet type, for the task of attribute classification. They show a large improvement on the task compared to HOG-based approaches. However, their approach includes a number of suboptimal choices. They use pre-trained HOG poselets to detect parts and they train a “shallow” CNN

(by today’s standards) from scratch using a relatively small dataset of 25k images. We train poselet-like part detectors on a much richer feature representation than HOG, derived from the pool5 layer of [18]. Indeed, [12, 13] show an impressive jump in object detection performance using pool5 instead of HOG. In addition, the task-specific CNN that we use for action or attribute classification shares the architecture of [18, 23] and is initialized by pre-training on the large ImageNet-1k dataset prior to task-specific fine-tuning.

In the same vein, Branson *et al.* [5] tackle the problem of bird species categorization by first detecting bird parts with a HOG-DPM and then extracting CNN features from the aligned parts. They experimentally show the superiority of CNN-based features to hand-crafted representations. However, they work from relatively weak HOG-DPM part detections, using CNNs solely for classification purposes. Switching to the person category, HOG-DPM does not generate accurate part/keypoint predictions as shown by [15], and thus cannot be regarded as a source for well aligned body parts.

**Deep parts.** Zhang *et al.* [31] introduce part-based R-CNNs for the task of bird species classification. They discover parts of birds from region proposals and combine them for classification. They gain from using parts and also from fine-tuning a CNN for the task starting from ImageNet weights. However, region proposals are not guaranteed to produce parts. Most techniques, such as [26], are designed to generate candidate regions that contain whole objects based on bottom-up cues. While this approach works for birds, it may fail in general as parts can be defined arbitrarily in an object and need not be of distinct color and texture with regard to the rest of the object. Our sliding-window parts provide a more general solution. Indeed, we find that the recall of selective search regions for our parts is 15.6% lower than our sliding-window parts across parts at 50% intersection-over-union.

Tompson *et al.* [24] and Chen and Yuille [6] train keypoint specific part detectors, in a CNN framework, for human body pose estimation and show significant improvement compared to [25]. Their models assume that all parts are visible or self-occluded, which is reasonable for the datasets they show results on. The data for our task contain significantly more clutter, truncation, and occlusion and so our system is designed to handle missing parts.

Bourdev *et al.* [4] introduce a form of deep poselets by training a network with a cross entropy loss. Their system uses a hybrid approach which first uses HOG poselets to bootstrap the collection of training data. They substitute deep poselets in the poselet detection pipeline [1] to create person hypotheses. Their network is smaller than [18] and they train it from scratch without hard negative mining. They show a marginal improvement over R-CNN for person detection, after feeding their hypothesis through R-

CNN for rescore and bounding box regression. Our parts look very much like poselets, since they capture parts of a pose. However, we cluster the space of poses instead of relying on random selection and train our models using a state-of-the-art network [18] with hard negative mining.

### 3. Deep part detectors

Figure 2 schematically outlines the design of our deep part detectors, which can be viewed as a multi-scale fully convolutional network. The first stage produces a feature pyramid by convolving the levels of the gaussian pyramid of the input image with a 5-layer CNN, similar to Girshick *et al.* [13] for training DeepPyramid DPMs. The second stage outputs a pyramid of part scores by convolving the feature pyramid with the part models.

#### 3.1. Feature pyramid

Feature pyramids allow for object and part detections at multiple scales while the corresponding models are designed at a single scale. This is one of the oldest “tricks” in computer vision and has been implemented by sliding-window object detection approaches throughout the years [1, 9, 22, 27].

Given an input image, the construction of the feature pyramid starts by creating the gaussian pyramid for the image for a fixed number of scales and subsequently extracting features from each scale. For feature extraction, we use a CNN and more precisely, we use a variant of the single-scale network proposed by Krizhevsky *et al.* [18]. More details can be found in [13]. Their software is publicly available and we build on their implementation.

#### 3.2. Part models

We design models to capture parts of the human body under a particular viewpoint and pose. Ideally, part models should be (a) pose-sensitive, *i.e.* produce strong activations on examples of similar pose and viewpoint, (b) inclusive, *i.e.* cover all the examples in the training set, and (c) discriminative, *i.e.* score higher on the object than on the background. To achieve all the above properties, we build part models by clustering the keypoint configurations of all the examples in the training set and train linear SVMs on pool5 features with hard negative mining.

##### 3.2.1 Designing parts

We model the human body with three high-level parts: the head, the torso and the legs. Even though the pose of the parts is tied with the global pose of the person, each one has its own degrees of freedom. In addition, there is a large, yet not infinite due to the kinematic constraints of the human body, number of possible part combinations that cover the space of possible human poses.

We design parts defined by the three body areas, head ( $H$ ), torso ( $T$ ) and legs ( $L$ ). Assume  $t \in \{H, T, L\}$  and  $K_t^{(i)}$  the set of 2D keypoints of the  $i$ -th training example corresponding to part  $t$ . The keypoints correspond to predefined landmarks of the human body. Specifically,  $K_H = \{Eyes, Nose, Shoulders\}$ ,  $K_T = \{Shoulders, Hips\}$  and for  $K_L = \{Hips, Knees, Ankles\}$ .

For each  $t$ , we cluster the set of  $K_t^{(i)}$ ,  $i = 1, \dots, N$ , where  $N$  is the size of the training set. The output is a set of clusters  $C_t = \{c_j\}_{j=1}^{P_t}$ , where  $P_t$  is the number of clusters for  $t$ , and correspond to distinct part configurations

$$C_t = \text{cluster} \left( \{K_t^{(i)}\}_{i=1}^N \right). \quad (1)$$

We use a greedy clustering algorithm, similar to [14]. Examples are processed in a random order. An example is added to an existing cluster if its distance to the center is less than  $\epsilon$ , otherwise it starts a new cluster. The distance of two examples is defined as the euclidean distance of their normalized keypoint distributions. For each cluster  $c \in C_t$ , we collect the  $M$  closest cluster members to its center. Those form the set of positive examples that represent the cluster. From now on, we describe a part by its body part type  $t$  and its cluster index  $j$ , with  $c_j \in C_t$ , while  $S_{t,j}$  represents the set of positive examples for part  $(t, j)$ .

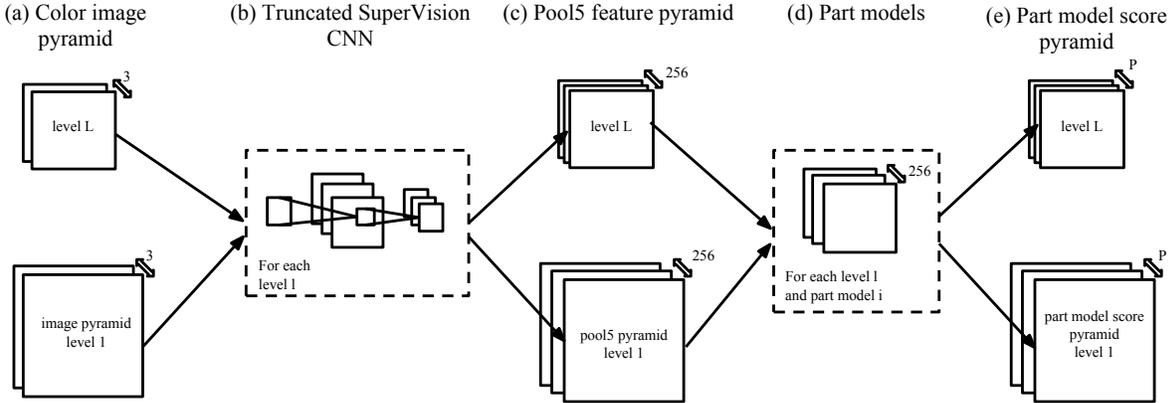
Figure 3 (left) shows examples of clusters as produced by our clustering algorithm with  $\epsilon = 1$  and  $M = 100$ . We show 4 examples for each cluster example. We use the PASCAL VOC 2012 train set, along with keypoint annotations as provided by [1], to design and train the part detectors. In total we obtain 30 parts, 13 for head, 11 for torso and 6 for legs.

##### 3.2.2 Learning part models

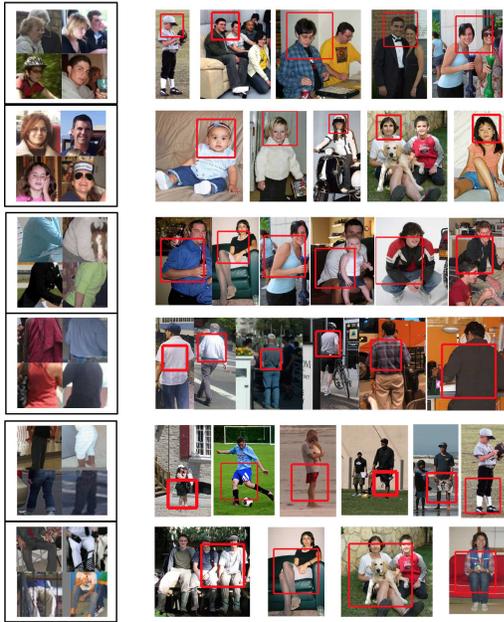
For each part  $(t, j)$ , we define the part model to be the vector of weights  $\mathbf{w}_{t,j}$  which when convolved with a feature pyramid gives stronger activations near the ground-truth location and scale (right most part of Figure 2).

One could view the whole pipeline shown in Figure 2 as a fully convolutional model and thus one could train it end-to-end, optimizing the weights of the CNN for the pool5 feature extraction and the weights of the part models jointly. We choose to simplify the problem by decoupling it. We use the publicly available ImageNet weights of the CNN [12] to extract pool5 feature pyramids. Subsequently, we train linear SVMs for the part models. For each part  $(t, j)$  we train a linear SVM with positives from  $S_{t,j}$  to obtain model weights  $\mathbf{w}_{t,j} \in \mathbb{R}^{8 \times 8 \times 256}$ . We use hard negative mining from images of no people to train the model.

Figure 3 (right) shows the top few detections of a subset of parts on PASCAL VOC val 2009 set. Each row shows activations of a different part, which is displayed at the left side of the same row.



**Figure 2:** Schematic overview of our part detectors. (a) A gaussian pyramid is build from an input image. (b) Each level of the pyramid is fed into a truncated SuperVision CNN. (c) The output is a pyramid of pool5 feature maps. (d) Each level of the feature pyramid is convolved with the part models. (e) The output is a pyramid of part model scores



**Figure 3:** Examples of clusters for the three body areas, head, torso and legs (left) and their top few detections on PASCAL VOC val 2009 (right). The first two rows correspond to cluster examples for head, the following two for torso and the last two for legs.

**Evaluation of part models.** We quantify the performance of our part detectors by computing the average precision (AP) - similar to object detection PASCAL VOC - on val 2009. For every image, we detect part activations at all scales and locations which we non-maximum suppress with a threshold of 0.3 across all parts of the same type. Since there are available keypoint annotations on the val set, we are able to construct ground-truth part boxes. A detection

AP (%)	$\sigma = 0.2$	$\sigma = 0.3$	$\sigma = 0.4$	$\sigma = 0.5$
Head	55.2	51.8	45.2	31.6
Torso	42.1	36.3	23.6	9.4
Legs	34.9	27.9	20.0	10.4

**Table 1:** AP for each part type on PASCAL VOC val 2009. We evaluate the part activations and measure AP for different thresholds of intersection-over-union.

is marked as positive if the intersection-over-union with a ground-truth part box is more than  $\sigma$ . In PASCAL VOC,  $\sigma$  is set to 0.5. However, this threshold is rather strict for small objects, such as our parts. We report AP for various values of  $\sigma$  for a fair assessment of the quality of our parts. Table 1 shows the results.

**Mapping parts to instances.** Since our part models operate independently, we need to group part activations and link them to an instance in question. Given a candidate region  $box$  in an image  $I$ , for each part  $t$  we keep the highest scoring part within  $box$

$$j^* = \operatorname{argmax}_j \max_{(x,y) \in box} \mathbf{w}_{t,j} * F_{(x,y)}(I), \quad (2)$$

where  $F_{(x,y)}(I)$  is the point in feature pyramid for  $I$  corresponding to the image coordinates  $(x, y)$ . This results in three parts being associated with each  $box$ , as shown in Figure 1. A part is considered absent if the score of the part activation is below a threshold, here the threshold is set to  $-0.1$ .

In the case when an oracle gives ground-truth bounding boxes at test time, one can refine the search of parts even further. If  $box$  is the oracle box in question, we retrieve the  $k$  nearest neighbor instances  $i = \{i_1, \dots, i_k\}$  from the train-

ing set based on the  $L_2$ -norm of their pool5 feature maps  $F(\cdot)$ , i.e.  $\frac{F(box)^T F(box_{i_j})}{\|F(box)\| \cdot \|F(box_{i_j})\|}$ . If  $K_{i_j}$  are the keypoints for the nearest examples, we consider the average keypoint locations  $K_{box} = \frac{1}{K} \sum_{j=1}^k K_{i_j}$  to be an estimate of the keypoints for the test instance  $box$ . Based on  $K_{box}$  we can reduce the regions of interest for each part within  $box$  by only searching for them in the corresponding estimates of the body parts.

## 4. Part-based Classification

In this section we investigate the role of parts for fine-grained classification tasks. We focus on the tasks of action classification (e.g. running, reading, etc.) and attribute classification (e.g. male, wears hat, etc.). Figure 4 schematically outlines our approach at test time. We start with the part activations mapped to an instance and forward propagate the corresponding part and instance boxes through a CNN. The output is a fc7 feature vector for each part as well as the whole instance. We concatenate the feature vectors and classify the example with a linear SVM, which predicts the confidence for each class (action or attribute).

### 4.1. System variations

For each task, we consider four variants of our approach in order to understand which design factors are important.

**No parts.** This approach is our baseline and does not use part detectors. Instead, each instance is classified according to the fc7 feature vector computed from the instance bounding box. The CNN used for this system is fine-tuned from an ImageNet initialization, as in [12], on jittered instance bounding boxes.

**Instance fine-tuning.** This method uses our part detectors. Each instance is classified based on concatenated fc7 feature vectors from the instance and all three parts. The CNN used for this system is fine-tuned on instances, just as in the “no parts” system. We note that because some instances are occluded, and due to jittering, training samples may resemble parts, though typically only the head and torso (since occlusion tends to happen from the torso down).

**Joint fine-tuning.** This method also uses our part detectors and concatenated fc7 feature vectors. However, unlike the previous two methods we fine-tune the CNN jointly using instance and part boxes from each training sample. During fine-tuning the network can be seen as a four-stream CNN, with one stream for each bounding box. Importantly, we tie weights between the streams so that the number of CNN parameters is the same in all system variants. This design explicitly forces the CNN to see each part box during fine-tuning.

**3-way split.** To test the importance of our part detectors, we employ a baseline that vertically splits the instance bounding box into three (top, middle, and bottom) in order to simulate crude part detectors. This variation uses a CNN fine-tuned on instances.

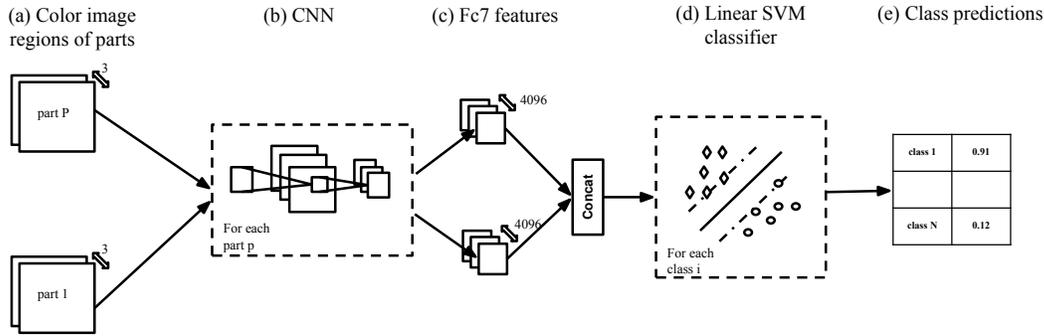
### 4.2. Action Classification

We focus on the problem of action classification as defined by the PASCAL VOC action challenge. The task involves predicting actions from a set of predefined action categories.

**Learning details.** We train all networks with backpropagation using Caffe [17], starting from the ImageNet weights, similar to the fine-tuning procedure introduced in [12]. A small learning rate of  $10^{-5}$  and a dropout ratio of 50% were used. During training, and at test time, if a part is absent from an instance then we use a box filled with the ImageNet mean image values (i.e., all zeros after mean subtraction). Subsequently, we train linear SVMs, one for each action, on the concatenated fc7 feature vectors.

**Context.** In order to make the most of the context in the image, we rescore our predictions by using the output of R-CNN [12] for the 20 PASCAL VOC object categories and the presence of other people performing actions. We train a linear SVM on the action score of the test instance, the maximum scores of other instances (if any) and the object scores, to obtain a final prediction. Context rescoring is used for all system variations on the *test* set.

**Results.** Table 2 shows the result of our approach on the PASCAL VOC 2012 *test* set. These results are in the standard setting, where an oracle gives ground-truth person bounds at test time. We conduct experiments using two different network architectures: a 8-layer CNN as defined in [18], and a 16-layer as defined in [23]. *Ours (no parts)* is the baseline approach, with no parts. *Ours* is our full approach when we include the parts. For the 8-layer network, we use the CNN trained on instances, while for the 16-layer network we use the CNN trained jointly on instances and their parts based on results on the *val* set (Table 3). For our final system, we also present results when we add features extracted from the whole image, using a 16-layer network trained on ImageNet-1k (*Ours (w/ image features)*). We show results as reported by action poselets [20], a part-based approach, using action specific poselets with HOG features, Oquab *et al.* [21], Hoai [16] and Simonyan and Zisserman [23], three CNN-based approaches on the task. The best performing method by [23] uses a 16- and 19-layer network. Their 16-layer network is equivalent to *Ours (no parts)* with 16 layers, thus the additional boost in perfor-



**Figure 4:** Schematic overview of our approach for fine grained classification using parts. (a) We consider regions of part activations. (b) Each part is forward propagated through a CNN. (c) The output is the fc7 feature vector for each input. (d) The features are concatenated and fed into linear SVM classifiers. (e) The classifiers produce scores for each class.

mance comes from the 19-layer network. This is not surprising, since deeper networks perform better, as is also evident from our experiments. From the comparison with the baseline, we conclude that parts improve the performance. For the 8-layer CNN, parts contribute 3% of mAP, with the biggest improvement coming from *Phoning*, *Reading* and *Taking Photo*. For the 16-layer CNN, the improvement from parts is smaller, 1.7 % of mAP, and the actions benefited the most are *Reading*, *Taking Photo* and *Using Computer*. The image features capture cues from the scene and give an additional boost to our final performance.

Table 3 shows results on the PASCAL VOC action *val* set for a variety of different implementations of our approach. *Ours (no parts)* is the baseline approach, with no parts, while *Ours (3-way split)* uses as parts the three horizontal splits comprising the instance box. *Ours (joint fine-tuning)* shows the results when using a CNN fine-tuned jointly on instances and parts, while *Ours (instance fine-tuning)* shows our approach when using a CNN fine-tuned on instances only. We note that all variations that use parts significantly outperform the *no-parts* system.

We also show results of our best system when ground-truth information is *not* available at test time *Ours (R-CNN bbox)*. In place of oracle boxes we use R-CNN detections for person. For evaluation purposes, we associate a R-CNN detection to a ground-truth instance as following: we pick the highest scoring detection for person that overlaps more than 0.5 with the ground truth. Another option would be to define object categories as “person+action” and then just follow the standard detection AP protocol. However, this is not possible because not all people are marked in the dataset (this is true for the attribute dataset as well). We report numbers on the val action dataset. We observe a drop in performance, as expected due to the imperfect person detector, but our method still works reasonably well under those circumstances. Figure 5 shows the top few predictions on the test

set. Each block corresponds to a different action.

### 4.3. Attribute Classification

We focus on the problem of attribute classification, as defined by [2]. There are 9 different categories of attributes, such as *Is Male*, *Has Long Hair*, and the task involves predicting attributes, given the location of the people. Our approach is shown in Figure 4. We use the Berkeley Attributes of People Dataset as proposed by [2].

**Learning details.** Similar to the task of action classification, we separately learn the parameters of the CNN and the linear SVM. Again, we fine-tune a CNN for the task in question with the difference that the softmax layer is replaced by a cross entropy layer (sum of logistic regressions).

**Results.** Table 4 shows AP on the *test* set. We show results of our approach with and without parts, as well as results as reported by Zhang *et al.* [33], the state-of-the-art on the task, on the same test set. With an 8-layer network, parts improve the performance of all categories, indicating their impact on attribute classification. Also, a network jointly fine-tuned on instances and parts seems to work significantly better than a CNN trained solely on the instance boxes. In the case of a 16-layer network, joint fine-tuning and instance fine-tuning seem to work equally well. The gain in performance from adding parts is less significant in this case. This might be because of the already high performance achieved by the holistic network. Interestingly, our 8-layer holistic approach matches the current state-of-the-art on this task, PANDA [33] showcasing the importance of deeper networks and good initialization.

Table 4 also shows the effectiveness of our best model, namely the jointly fine-tuned 16-layer CNN, when we use R-CNN detections instead of ground truth boxes on the Berkeley Attributes of People test set. Figure 7 shows the

AP (%)	CNN layers	Jumping	Phoning	Playing Instrument	Reading	Riding Bike	Riding Horse	Running	Taking Photo	Using Computer	Walking	mAP
Action Poselets [20]	-	59.3	32.4	45.4	27.5	84.5	88.3	77.2	31.2	47.4	58.2	55.1
Oquab <i>et al.</i> [21]	8	74.8	46.0	75.6	45.3	93.5	95.0	86.5	49.3	66.7	69.5	70.2
Hoai [16]	8	82.3	52.9	84.3	53.6	95.6	96.1	89.7	60.4	76.0	<b>72.9</b>	76.3
Simonyan & Zisserman [23]	16 & 19	<b>89.3</b>	<b>71.3</b>	<b>94.7</b>	<b>71.3</b>	<b>97.1</b>	<b>98.2</b>	<b>90.2</b>	73.3	<b>88.5</b>	66.4	<b>84.0</b>
Ours (no parts)	8	76.2	47.4	77.5	42.2	94.9	94.3	87.0	52.9	66.5	66.5	70.5
Ours	8	77.9	54.5	79.8	48.9	95.3	95.0	86.9	61.0	68.9	67.3	73.6
Ours (no parts)	16	84.7	62.5	86.6	59.0	95.9	96.1	88.7	69.5	77.2	70.2	79.0
Ours	16	83.7	63.3	87.8	64.2	96.0	96.7	88.9	75.2	80.0	71.5	80.7
Ours (w/ image features)	16	84.7	67.8	91.0	66.6	96.6	97.2	<b>90.2</b>	<b>76.0</b>	83.4	71.6	82.6

**Table 2:** AP on the PASCAL VOC 2012 Actions test set. The first three rows show results of two other methods. Action Poselets [20] is a part based approach using HOG features, while Oquab *et al.* [21], Hoai [16] and Simonyan & Zisserman [23] are CNN based approaches. *Ours (no parts)* is the baseline approach of our method, when only the ground truth box is considered, while *Ours* is the full approach, including parts. All approaches use ground truth boxes at test time.

AP (%)	CNN layers	Jumping	Phoning	Playing Instrument	Reading	Riding Bike	Riding Horse	Running	Taking Photo	Using Computer	Walking	mAP
Ours (no parts)	8	76.5	44.7	75.0	43.3	90.0	91.6	79.2	53.5	66.5	61.4	68.2
Ours (3-way split)	8	79.0	44.4	77.8	46.9	91.5	93.1	83.1	59.3	67.3	64.4	70.7
Ours (instance fine-tuning)	8	77.2	48.4	79.0	49.5	92.4	93.8	80.9	60.4	68.9	64.0	71.5
Ours (joint fine-tuning)	8	75.2	49.5	79.5	50.2	93.7	93.6	81.5	58.6	64.6	63.6	71.0
Ours (no parts)	16	<b>85.4</b>	58.6	84.6	60.9	94.4	96.6	<b>86.6</b>	68.7	74.9	67.3	77.8
Ours (instance fine-tuning)	16	85.1	60.2	<b>86.6</b>	63.1	95.6	97.4	86.4	71.0	77.6	68.3	79.1
Ours (joint fine-tuning)	16	84.5	<b>61.2</b>	88.4	<b>66.7</b>	<b>96.1</b>	<b>98.3</b>	85.7	<b>74.7</b>	<b>79.5</b>	<b>69.1</b>	<b>80.4</b>
Ours (R-CNN bbox)	8	67.8	46.6	76.9	47.3	85.9	81.4	71.5	53.1	61.2	53.9	64.6
Ours (R-CNN box)	16	79.4	63.3	86.1	64.4	93.2	91.9	80.2	71.2	77.4	63.4	77.0

**Table 3:** AP on the PASCAL VOC 2012 Actions val set of our approach. *Ours (no parts)* is our approach without parts. *Ours (3-way split)* is our approach when parts are defined as the three horizontal splits comprising an instance box. *Ours (joint fine-tuning)* uses a CNN fine-tuned jointly on the instances and the parts, while *Ours (instance fine-tuning)* uses a single CNN fine-tuned just on the instance box. All the above variations of our approach use ground truth information at test time as the object bound. *Ours (R-CNN bbox)* uses R-CNN detections for person.

AP (%)	CNN layers	Is Male	Has Long Hair	Has Glasses	Has Hat	Has T-Shirt	Has Long Sleeves	Has Shorts	Has Jeans	Has Long Pants	mAP
PANDA [33]	5	91.7	82.7	70.0	74.2	49.8	86.0	79.1	81.0	96.4	79.0
Ours (no parts)	8	87.5	80.4	43.3	77.0	61.5	86.4	88.5	88.7	98.2	79.1
Ours (3-way split)	8	89.3	82.2	51.2	84.0	60.1	87.4	88.3	89.2	98.2	81.1
Ours (instance fine-tuning)	8	89.9	83.5	60.5	85.2	64.3	89.0	88.6	89.1	98.2	83.1
Ours (joint fine-tuning)	8	91.7	86.3	72.5	89.9	69.0	90.1	88.5	88.3	98.1	86.0
Ours (no parts)	16	93.4	88.7	72.5	91.9	72.1	94.1	92.3	91.9	98.8	88.4
Ours (instance fine-tuning)	16	<b>93.8</b>	89.8	76.2	92.9	<b>73.3</b>	<b>94.4</b>	92.3	91.8	98.7	89.3
Ours (joint fine-tuning)	16	92.9	<b>90.1</b>	<b>77.7</b>	<b>93.6</b>	72.6	93.2	<b>93.9</b>	<b>92.1</b>	<b>98.8</b>	<b>89.5</b>
Ours (R-CNN bbox)	8	84.1	77.9	62.7	84.5	66.8	84.7	80.7	79.2	91.9	79.2
Ours (R-CNN bbox)	16	90.1	85.2	70.2	89.8	63.2	89.7	83.4	84.8	96.3	83.6

**Table 4:** AP on the test set of the Berkeley Attributes of People Dataset. All approaches on the top use ground truth boxes for evaluation. *Ours (no parts)* is the baseline approach with no parts. *Ours (3-way split)* is a variant of our approach, where parts are defined as the three horizontal splits comprising an instance box. *Ours (instance fine-tuning)* uses a CNN fine-tuned on instance boxes, while *Ours (joint fine-tuning)* uses a CNN fine-tuned jointly on instances and parts. We also show the effectiveness of our approach *Ours (R-CNN bbox)*, when no ground truth boxes are given at test time.

top few predictions on the test set. Each block corresponds to a different attribute. Figure 6 shows top errors for two of our lowest performing attribute classes.

## Acknowledgements

This work was supported by the Intel Visual Computing Center and the ONR SMARTS MURI N000140911051. The GPUs used in this research were generously donated by the NVIDIA Corporation



**Figure 6:** Top errors of classification for two of the attribute categories, *Has Glasses* (top) and *Has T-Shirt* (bottom).

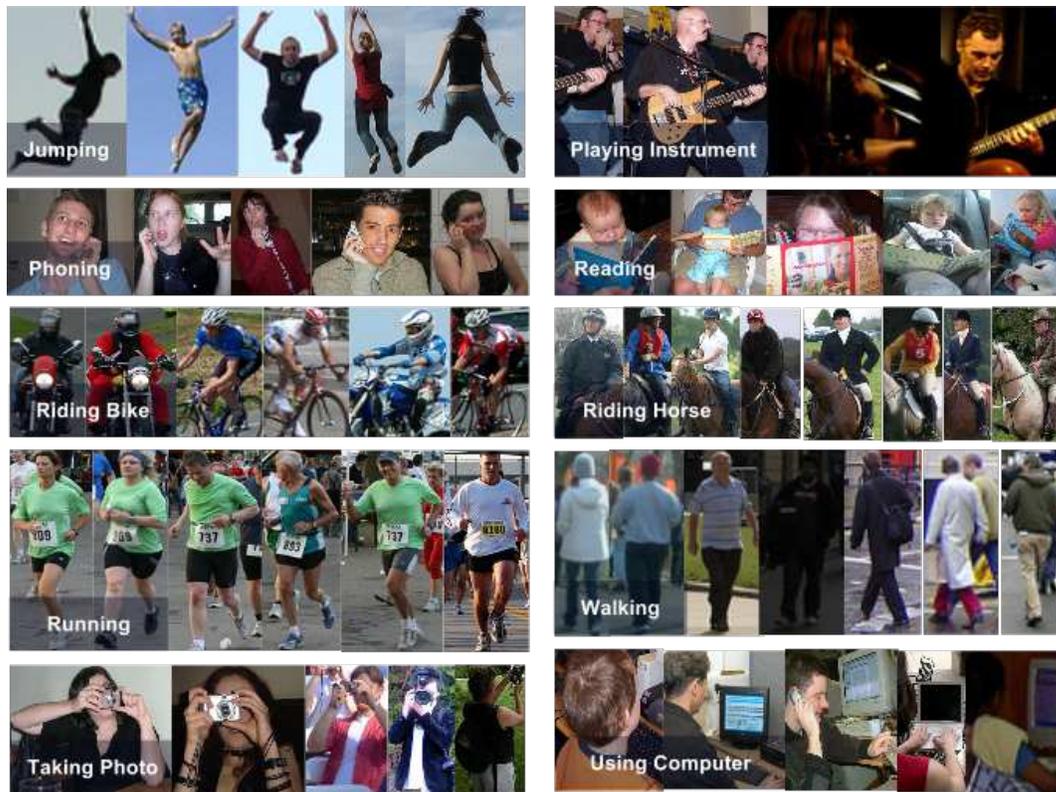


Figure 5: Top action predictions on the test set. Different blocks correspond to different actions.



Figure 7: Top attribute predictions on the test set. Each block corresponds to a different attribute

## References

- [1] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *ECCV*, 2010. 1, 2, 3
- [2] L. Bourdev, S. Maji, and J. Malik. Describing people: Poselet-based attribute classification. In *ICCV*, 2011. 2, 6
- [3] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009. 1, 2
- [4] L. Bourdev, F. Yang, and R. Fergus. Deep poselets for human detection. *arXiv preprint arXiv:1407.0717*, 2014. 2
- [5] S. Branson, G. V. Horn, S. Belongie, and P. Perona. Bird species categorization using pose normalized deep convolutional nets. In *BMVC*, 2014. 2
- [6] X. Chen and A. Yuille. Articulated pose estimation with image-dependent preference on pairwise relations. In *NIPS*, 2014. 2
- [7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 1
- [8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes (VOC) Challenge. *IJCV*, 2010. 2
- [9] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *TPAMI*, 2010. 2, 3
- [10] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 2005. 2
- [11] M. Fischler and R. Elschlager. The representation and matching of pictorial structures. *IEEE Trans. on Computer*, 22(1), 1973. 2
- [12] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 2, 3, 5
- [13] R. Girshick, F. Iandola, T. Darrell, and J. Malik. Deformable part models are convolutional neural networks. *arXiv preprint arXiv:1409.5403*, 2014. 1, 2, 3
- [14] G. Gkioxari, P. Arbelaez, L. Bourdev, and J. Malik. Articulated pose estimation using discriminative armlet classifiers. In *CVPR*, 2013. 2, 3
- [15] G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik. Using k-poselets for detecting people and localizing their keypoints. In *CVPR*, 2014. 2
- [16] M. Hoai. Regularized max pooling for image categorization. In *BMVC*, 2014. 5, 7
- [17] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014. 5
- [18] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012. 2, 3, 5
- [19] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1989. 2
- [20] S. Maji, L. Bourdev, and J. Malik. Action recognition from a distributed representation of pose and appearance. In *CVPR*, 2011. 1, 2, 5, 7
- [21] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 2014. 1, 5, 7
- [22] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *ICLR*, 2014. 3
- [23] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 2, 5, 7
- [24] J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS*, 2014. 2
- [25] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, 2014. 2
- [26] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. *IJCV*, 2013. 2
- [27] R. Vaillant, C. Monrocq, and Y. L. Cun. An original approach for the localization of objects in images. *IEE Proc on Vision, Image, and Signal Processing*, 1994. 3
- [28] Y. Wang, D. Tran, and Z. Liao. Learning hierarchical poselets for human parsing. In *CVPR*. IEEE, 2011. 2
- [29] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures-of-parts. *TPAMI*, 2012. 2
- [30] B. Yao, A. Khosla, and L. Fei-Fei. Combining randomization and discrimination for fine-grained image categorization. In *CVPR*, 2011. 1
- [31] N. Zhang, J. Donahue, R. Girshick, and T. Darrell. Part-based R-CNNs for fine-grained category detection. In *ECCV*, 2014. 1, 2
- [32] N. Zhang, R. Farrell, F. Iandola, and T. Darrell. Deformable part descriptors for fine-grained recognition and attribute prediction. In *ICC*, 2013. 2
- [33] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev. PANDA: Pose aligned networks for deep attribute modeling. In *CVPR*, 2014. 1, 2, 6, 7