Joint Multi-Person Pose Estimation and Semantic Part Segmentation

Fangting Xia[1] sukixia@gmail.com
Peng Wang[1] pengwangpku2012@gmail.com
Xianjie Chen[1] cxj@ucla.edu
Alan Yuille[2] alan.yuille@jhu.edu

[1]University of California, Los Angeles
Los Angeles, CA 90095

[2]Johns Hopkins University
Baltimore, MD 21218

Abstract

Human pose estimation and semantic part segmentation are two complementary tasks in computer vision. In this paper, we propose to solve the two tasks jointly for natural multi-person images, in which the estimated pose provides object-level shape prior to regularize part segments while the part-level segments constrain the variation of pose locations. Specifically, we first train two fully convolutional neural networks (FCNs), namely Pose FCN and Part FCN, to provide initial estimation of pose joint potential and semantic part potential. Then, to refine pose joint location, the two types of potentials are fused with a fully-connected conditional random field (FCRF), where a novel segment-joint smoothness term is used to encourage semantic and spatial consistency between parts and joints. To refine part segments, the refined pose and the original part potential are integrated through a Part FCN, where the skeleton feature from pose serves as additional regularization cues for part segments. Finally, to reduce the complexity of the FCRF, we induce human detection boxes and infer the graph inside each box, making the inference forty times faster.

Since there's no dataset that contains both part segments and pose labels, we extend the PASCAL VOC part dataset [6] with human pose joints[1] and perform extensive experiments to compare our method against several most recent strategies. We show that our algorithm surpasses competing methods by 10.6% in pose estimation with much faster speed and by 1.5% in semantic part segmentation.

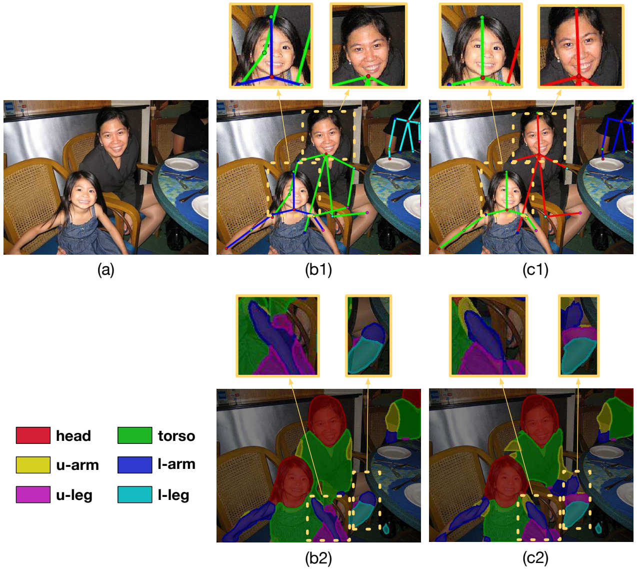

Figure 1: Joint human pose estimation and semantic part segmentation improve both tasks. (a) input image. (b) pose estimation and semantic part segmentation results before joint inference. (c) pose estimation and semantic part segmentation results after joint inference. Note that comparing (b1) and (c1), our result recovers the missing forehead joint and corrects the location error of right elbow and right wrist for the woman on the right. Comparing (b2) and (c2), our result gives more accurate details of lower arms and upper legs than (b2) for both people.

1. Introduction

Human pose estimation (*i.e*. predicting the position of joints for each human instance) and semantic part segmentation (*i.e*. decomposing humans into semantic part regions) are two crucial and correlated tasks in analysing humans from images. They provide richer representations for many dependent tasks, *e.g*. fine-grained recognition [1, 38, 17], action recognition [32, 30], image/video retrieval [36, 16], person-identification [24] and video surveillance [23].

Recently, dramatic progress has been made on pose estimation [8, 7, 34, 25] and human part segmentation [3, 31, 33, 20] with the advent of powerful convolutional neural networks (CNN) [19] and the availability of pose/segment annotations on large-scale datasets [12, 6, 21]. However,

[1] https://sukixia.github.io/paper.html

the two tasks are mostly solved independently without considering their correlations. As shown in the middle column in Fig. 1, for pose estimation, by designing loss w.r.t. the joints solely, it may omit the knowledge of dense pixel-wise part appearance coherence, yielding joints located outside of human instance or misleading joints when two people are close to each other. On the other hand, for part segmentation, through training that only respects pixel-wise part labels, it lacks proper overall human shape regularization, yielding missing/errorneous predictions when appearance cues are weak or missing.

In fact, the two tasks are complementary, and solving them jointly can reduce the learning difficulty in addressing each of them individually. As shown in the right column Fig. 1, by handling the two tasks jointly, the ambiguity in pose estimation (*e.g.* out of instance region) can be corrected by considering semantic part segments, while the estimated pose skeleton provides object-level context and regularity to help part segments align with human instances, *e.g.* over the details of arms and legs where appearance cues are missing.

Specifically, we illustrate our framework in Fig. 2. Firstly, given an image that contains multiple people, we train two FCNs: Pose FCN and Part FCN. Similar to [15], the Pose FCN outputs the pixel-wise joint score map, *i.e.* the potential of joints at each pixel (how likely a type of joint is located at certain pixel), and also outputs the joint neighbour score map, *i.e.* the potential of the location likelihood of neighboring joints for each joint type. The Part FCN produces the part score map for each semantic part type. Secondly, the three types of information are fused through a FCRF to refine the human joint locations, where a novel smoothness term on both part segments and joint proposals (generated from the initially estimated pixel-wise joint score map) are applied to encourage the consistency between segments and joints. Thirdly, the refined pose joints are re-organized into pose features that encode overall shape information, and are fed into a second-stage Part FCN as an additional input besides the initial part score map, yielding better segmentation results. To reduce the complexity of the FCRF, rather than infer over the full image as [15], we adopt a human detector [26] to first get the bounding box for each human instance and resize each instance region in a similar way to [33]. Our whole inference procedure is then performed within each resized region.

Last but not the least, in order to train and evaluate our method, we augment the challenging PASCAL-Person-Part dataset [6] with 14 human pose joint locations through manual labeling and make the annotations public. This dataset includes 3533 images that contain large variation of human poses, scales and occlusion. We evaluate our method over this dataset, and show that our approach outperforms the most recent competing methods for both tasks. In particular, our method is more effective and much faster (8 seconds versus 4 minutes) than DeeperCut [15] which is arguably the most effective algorithm for multi-person pose estimation.

In summary, the contributions of this paper lay in three folds: (1) to our best knowledge, we are the first to explore and demonstrate the complementary property of multi-person pose estimation and part segmentation with deep learned potentials; (2) by combining detection boxes in the pipeline, we reduce the complexity of FCRF inference over the full image, yielding better efficiency; (3) we extend the well labelled PASCAL-Person-Part dataset with human joints and demonstrate the effectiveness of our approach.

2. Related Works

Pose estimation. Traditional approaches use graphical models to combine spatial constraints with local observations of joints, based on low-level features [13, 37]. With the growing popularity of deep learning, recent methods rely on strong joint detectors trained by DCNNs [8, 28], and often use a simple graphical model (*e.g.* tree model, And-Or graph) to select and assemble joints into a valid pose configuration. These recent methods perform much better than traditional ones, but the localization of joints is still inaccurate (*e.g.* sometimes outside the human body) and they still struggle when there are multiple people overlapping each other. Other approaches discard graphical models by modeling the spatial dependencies of joints within DCNNs [29, 2, 9]. These approaches perform well on relatively simple datasets, but their ability to handle large pose variations in natural multi-person datasets is limited. A very recent work, Deeper-Cut [15], addresses the multi-person issue explicitly, using integer linear programming to cluster joint candidates into multiple human instances and assign joint types to each joint candidate. Deeper-Cut handles multi-person overlapping well, but is very time-consuming (4 minutes per image) and its performance on datasets with large scale variation is not fully satisfactory. Our method improves in these aspects by introducing a segment-joint consistency term that yields better localization of flexible joints such as wrists and ankles, and an effective scale handling strategy (using detected boxes and smart box rescaling) that can deal with humans of different sizes.

Semantic part segmentation. Previous approaches either use graphical models to select and assemble region proposals [34], or use fully convolutional neural networks (FCNs) [22] to directly produce pixel-wise part labels. Traditional graphical models [35, 11] find it difficult to handle the large variability of pose and occlusion in natural images. FCN-type approaches [3, 31], though simple and fast, give coarse part details due to FCN's inherent invariance property, and can have local confusion errors (*e.g.* la-

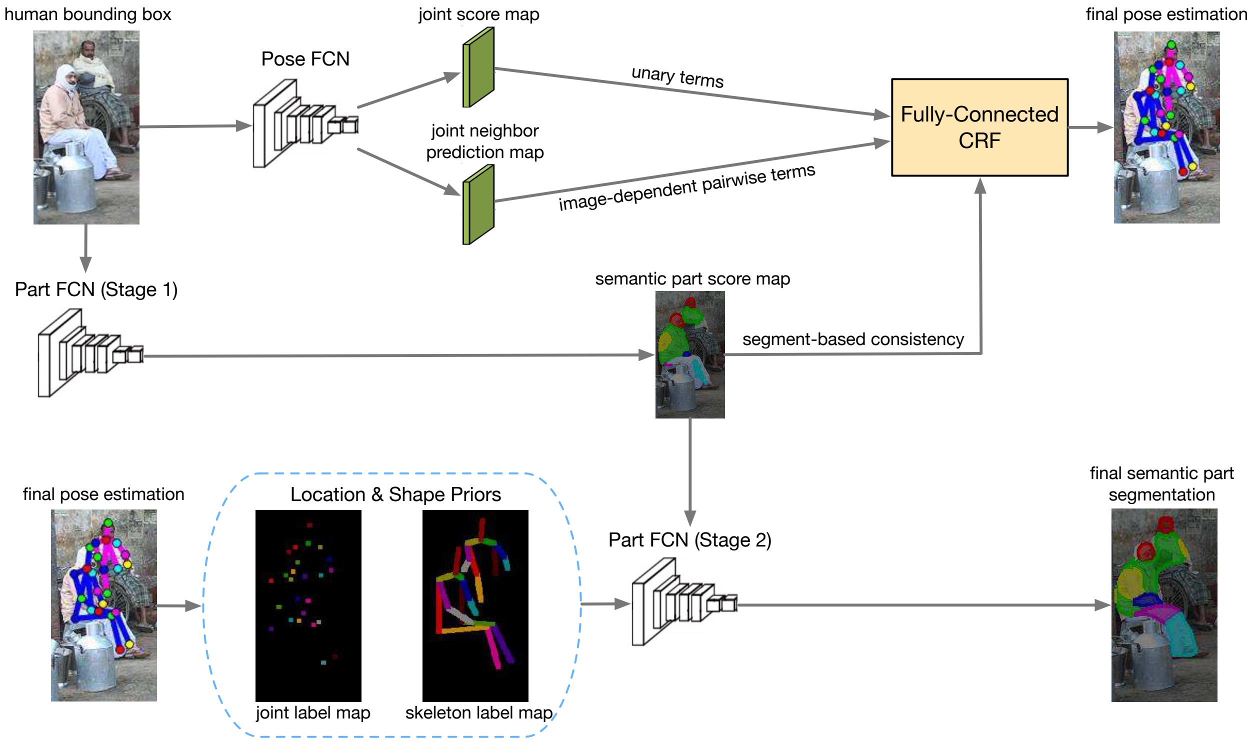

Figure 2: The framework of our approach for joint pose estimation and part segmentation. Initial joint scores and part segment scores are fused to yield better pose estimation results, and then the estimated poses are used to refine part segmentation.

beling arms as legs, labeling background regions as arms, *etc*.) if the person is in a non-typical pose, or when there are some other object/person nearby with similar appearance. Two recent works improve on FCN-type approaches by paying attention to the large scale variation in natural images. Chen *et al*. learn pixel-wise weights through an attention model [5] to combine the part segmentation results of three fixed scales. Xia *et al*. build an hierarchical model that adapts to object scales and part scales using "auto-zoom" [33]. We treat these two methods as our baselines, and demonstrate the advantages of our part segmentation approach. Most recently, researchers design and adopt more powerful network architectures such as Graph Long Short-Term Memory (LSTM) [20] and DeepLab with Deep Residual Net [4], greatly improving the performance. We prove that our method is complementary and can be added to these networks to further improve the performance.

Joint pose estimation and part segmentation. Yamaguchi *et al*. perform pose estimation and semantic part segmentation sequentially for clothes parsing, using a CRF with low-level features [35]. Ladicky *et al*. combine the two tasks in one principled formulation, using also low-level features [18]. Dong *et al*. combine the two tasks with a manually designed And-Or graph [10]. These methods demonstrate the complementary properties of the two tasks on relatively simple datasets, but they cannot deal with images with large pose variations or multi-person overlapping, mainly due to the less powerful features they use or the poor quality of their part region proposals. In contrast, our model combines FCNs with graphical models, greatly boosting the representation power of models to handle large pose variation. We also introduce novel part segment consistency terms for pose estimation and novel pose consistency terms for part segmentation, further improving the performance.

3. Our Approach

Given an image $\mathbf{I}$ with size h×w, our task is to output a pixel-wise part segmentation map $\mathbf{L}_s$, and a list of scored pose configurations $\mathcal{C}_p = \{(\mathbf{c}_i, s_i)|i = 1, 2, \ldots, k_i\}$, where $\mathbf{c}_i$ is the location of all 14 pose joint types for the person and s_i is the score of this pose configuration.

As illustrated in Fig. 2, for each human detection box, we first use Pose FCN and Part FCN to give initial estimation of pose location and part segmentation. Then a FCRF is used to refine pose estimation and a second-stage Part FCN is adopted for part refinement. Specifically, we first extract human bounding boxes with Faster R-CNN [26], and resize the image region within each detection box following [33] so that small people are enlarged and extra large people are shrunk to a fixed size. The resized box regions serve as input to Pose FCN and Part FCN. Pose FCN adopts the network architecture of ResNet-101 proposed in [14], while for Part FCN we use DeepLab-LargeFOV [3].

Pose FCN outputs two feature maps: (1) the pixel-wise joint score map $\mathbf{P}_j$, which is a matrix with shape h×w×14 representing the probability of each joint type locating at each pixel. (2) the pixel-wise joint neighbor score map $\mathbf{P}_n$, which is a h×w×364 matrix representing the probability of expected neighbor location for each joint. Here, the dimension of 364 is obtained by 14×13×2, which means for each joint the we estimate the other 13 joint locations using the offset $(\delta x, \delta y)$. Following the definition of parts in [3], Part FCN outputs a part score map $\mathbf{P}_s$ including 7 classes: 6 part labels and 1 background label.

Given the three score maps, we design a novel segment-joint smoothness term for our FCRF to obtain refined pose estimation results (detailed in Sec. 3.1). To obtain better part segmentation results, we further design a second-stage Part FCN, which takes joint input of first-stage part scores and derived feature maps from refined poses (detailed in Sec. 3.2). Finally, the estimated poses from each bounding box are merged through a Non-Maximum Suppression (NMS) strategy detailed in Sec. 4.1. For part segmentation, we merge the segment score map from different boxes using score averaging similar to [33].

3.1. Human Pose Estimation

In this section, we explain how we unify the three score maps (*i.e.* $\mathbf{P}_j$, $\mathbf{P}_n$ and $\mathbf{P}_s$) to estimate poses in each human detection box.

Following DeeperCut [15], we adopt a FCRF to obtain robust context for assembling the proposed joints into human instances. To reduce the complexity of the FCRF, rather than consider all the pixels, we generate 6 candidate locations for each joint from the joint score map $\mathbf{P}_j$ by non-maximum suppression (NMS). Formally, the FCRF for the graph is formulated as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where the node set $\mathcal{V} = \{c_1, c_2, \ldots, c_n\}$ represents all the candidate locations of joints and the edge set $\mathcal{E} = \{(c_i, c_j) | i = 1, 2, \ldots, n, j = 1, 2, \ldots, n, i < j\}$ is the edges connecting all the locations. The label to predict for each node is its joint type $l_{c_i} \in \{0, \cdots, K\}$, where $K = 14$ is the number of joint types and type 0 represents that the node belongs to background and is not selected. Besides, we also predict whether two nodes belong to the same person, *i.e.* $l_{c_i,c_j} \in \{0, 1\}$, where 1 indicates the two nodes are for the same person. Let $\mathcal{L} = \{l_{c_i} | c_i \in \mathcal{V}\} \cup \{l_{c_i,c_j} | (c_i, c_j) \in \mathcal{E}\}$. The target we want to optimize is:

$$\min_{\mathcal{L}} \sum_{c_i \in \mathcal{V}} \psi_i(l_{c_i}) + \sum_{(c_i,c_j) \in \mathcal{E}} \psi_{i,j}(l_{c_i}, l_{c_j}, l_{c_i,c_j}) \qquad (1)$$

where the unary term is defined as $\psi_i = \log \frac{1 - \mathbf{P}_j(l_{c_i})}{\mathbf{P}_j(l_{c_i})}$, which is a log-likelihood at location c_i based on the Pose-CNN output, the joint score map $\mathbf{P}_j$.

In contrast, the pairwise term is determined by both the joint neighbor score map $\mathbf{P}_n$ and the segmentation score map $\mathbf{P}_s$. Formally,

$$\psi_{i,j} = l_{c_i,c_j} \log \frac{1 - \mathbf{P}_{i,j}(l_{c_i}, l_{c_j} | \mathbf{P}_n, \mathbf{P}_s)}{\mathbf{P}_{i,j}(l_{c_i}, l_{c_j} | \mathbf{P}_n, \mathbf{P}_s)} \qquad (2)$$

where $\mathbf{P}_{i,j}(l_{c_i}, l_{c_j}) = \frac{1}{1+\exp(-\boldsymbol{\omega} \cdot \mathbf{f}(c_i, c_j, l_{c_i}, l_{c_j}))}$, obtained from logistic regression results w.r.t. a combined feature vector $\mathbf{f}$ from $\mathbf{f}(\mathbf{P}_n)$ and $\mathbf{f}(\mathbf{P}_s)$, in which we omit $c_i, c_j, l_{c_i}, l_{c_j}$ for simplicity.

The feature vector $\mathbf{f}(\mathbf{P}_n)$ encodes information to help decide whether the two proposals belong to the same person. We borrow the idea proposed in [15], and here we explain how the feature is extracted for paper completeness. Given the location of two joint proposals c_i, c_j and their corresponding label l_{c_i}, l_{c_j}, we first derive a direct vector from c_i to c_j, denoted as $\mathbf{v}_{i,j}$. In addition, given c_i, l_{c_i}, l_{c_j}, based on the joint neighbor offset score map $\mathbf{P}_n$, we may find an estimated location of l_{c_j} respecting c_i though computing $c'_j = c_i + (\delta x, \delta y)_{i,j}$. We denote the direct vector from c_i to the estimated location as $\mathbf{v}'_{i,j}$. Similar vectors $\mathbf{v}_{j,i}, \mathbf{v}'_{j,i}$ can be extracted in the same way. Feature $\mathbf{f}(\mathbf{P}_n) = [\ |\mathbf{v}_{j,i} - \mathbf{v}'_{j,i}|, |\mathbf{v}_{i,j} - \mathbf{v}'_{i,j}|, < \mathbf{v}_{j,i}, \mathbf{v}'_{j,i} >, < \mathbf{v}_{i,j}, \mathbf{v}'_{i,j} >\ ]$, in which $|. - .|$ is the euclidean distance between two vectors and $< . , . >$ is the angle between two vectors.

The feature vector $\mathbf{f}(\mathbf{P}_s)$ considers the correlation between joints and segments. Intuitively, joints are the connection points of parts. If two joints are neighboring joints, using forehead and neck as an example, the head joint should be located inside the head segment region and near the head segment boundary while the neck joint should be located in either head or body region and near the common boundary of body and head. Moreover, the connected line between forehead joint and neck joint should fall inside the head region. These segment-based heuristic cues provide strong constrains for the location of joints. We design $\mathbf{f}(\mathbf{P}_s)$ w.r.t. this idea. Formally, each joint type is associated with one or two semantic parts and each neighbouring joint type pair is associated with one semantic part type.

Based on the part segmentation label map $\mathbf{L}_s$ inferred from $\mathbf{P}_s$, here we introduce the feature $\mathbf{f}(\mathbf{P}_s)$ using the example of forehead and neck. For details, please see the supplementary material. Suppose l_{c_i}= forehead and l_{c_j}= neck, then our feature from segment includes 4 components: (1) a 2-d binary feature, with the first dimension indicating whether c_i is inside the head region, and the second dimension indicating whether it is around the boundary of the head region; (2) a 4-d binary feature, with the first 2-d feature indicating c_j w.r.t. the head region same as (1), and the rest 2-d feature indicating c_j w.r.t. the torso region respectively; (3) a 1-d feature indicating the proportion of pixels on the line segment between c_i and c_j that fall inside the head region; (4) a 1-d feature indicating the intersect-over-union (IOU) between an oriented rectangle computed from c_i and c_j (with aspect ratio = 2.5:1) and the head region. We only extract the full feature for neighboring joints. For the joints locating far away like head and feet, we drop the third and the fourth components of the feature and set them to be 0. We validate the parameters for aspect ratio through a mean human shape following [27].

Based on the unary and pairwise terms described above, the FCRF infers the best labels $\mathcal{L}$ for the generated joint proposals $c_1, c_2, \ldots, c_n$, selecting and assembling them into a list of pose configurations. We adapt the inference algorithm introduced in [15], transforming the FCRF into an integer linear programming (ILP) problem with additional constraints from $\mathcal{L}$. For each detection box, the inference algorithm gives the labels $\mathcal{L}$ for joint proposals within 1 sec. and we can acquire a list of pose configurations based on $\mathcal{L}$, with pose score equal to the sum of unary scores for all visible joints. For each detection box, we choose only one pose configuration whose center is closest to the detection box center, and add that pose configuration to our final pose estimation result. We also experiment with the strategy of extracting multiple pose configurations from each detection box since there might be multiple people in the detection box, but find this strategy doesn't improve the results.

3.2. Semantic Part Segmentation

We train a part segmentation model (the second-stage Part FCN) to segment an image into semantic parts with estimated high-quality pose configurations $\mathcal{C}_p$. We define two pose feature maps from $\mathcal{C}_p$: a joint label map and a skeleton label map, and use them as inputs to the second-stage Part FCN in addition to the original part score map. For the joint label map, we draw a circle with radius 3 at each joint location in $\mathcal{C}_p$. For the skeleton label map, we draw a stick with width 7 between neighbouring joints in $\mathcal{C}_p$. Fig. 2 illustrates the two simple and intuitive feature maps.

The second-stage Part FCN is much lighter than the first-stage Part FCN since we already have the part score map $\mathbf{P}_s$ predicted. We concatenate the 2 dimension feature map from estimated poses with the original part score map, yielding a 7 + 2 dimension inputs, and stacked 3 additional convolutional layers with kernel size as 7, kernel dimension as 128 and Relu as activation function. Our final part segmentation is then derived using the argmax value from the output part score map.

To learn all the parameters, we adopt a stage-wise strategy, *i.e.* first learn Pose FCN and the first-stage Part FCN, then the FCRF, and finally the second-stage Part FCN, which roughly take 3 days to train. For inference, our framework takes roughly 6s per-image. It is possible for us to do learning and inference iteratively. However, we found it's practically inefficient and the performance improvement is marginal. Thus, we only do the refinement once.

4. Experiments

4.1. Implementation Details

Data. We perform extensive experiments on our manually labeled dataset, PASCAL-Person-Part [6], which provides joint and part segment annotations for PASCAL person images with large variation in pose and scale. There are 14 annotated joint types (*i.e.* forehead, neck, left/right shoulder, l/r elbow, l/r wrist, l/r waist, l/r knee and l/r ankle) and we combine the part labels into 6 semantic part types (*i.e.* head, torso, upper arm, lower arm, upper leg and lower leg). We only use those images containing humans for training (1716 images) and validation (1817 images). We only experiment on this dataset because other datasets do not have both pose and part segment annotations.

Generation of joint proposals. We apply the Faster R-CNN detector to produce human detection boxes, and perform a NMS procedure with detection score threshold $= 0.6$ and box IOU overlap threshold = 0.6. For each human detection box, we generate 6 joint proposals per joint type from the joint score map outputted by Pose FCN, using a NMS procedure with joint score threshold $= 0.2$ and proposal distance threshold $= 16$.

Generation of final pose configurations. For each detection box, the FCRF selects and assembles joint proposals into a series of pose configurations, with pose score defined as the sum of all unary joint scores (in logarithm form). For each missing joint, we regard its unary score as 0.2. To combine pose configurations from all the detection boxes, we design a NMS prodedure which considers the overlap of head bounding box, upper-body bounding box, lower-body bounding box and whole-body bounding box inferred from the pose configurations. For two pose configurations, the one with a lower pose score will be filtered if their IOU overlap exceeds 0.65 for head boxes, or 0.5 for upper-body/lower-body boxes, or 0.4 for whole-body boxes.

4.2. Human Pose Estimation

Previous evaluation metrics (*e.g.* PCK and PCP) do not penalize false positives that are not part of the groundtruth. So following [15], we compare our model with other state-of-the-arts by Mean Average Precision (mAP). Briefly speaking, pose configurations in C_I^{pose} are first matched to groundtruth pose configurations according to the pose box overlap, and then the AP for each joint type is computed and

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	U-Body	Total (mAP)
Chen & Yuille	45.3	34.6	24.8	21.7	9.8	8.6	7.7	31.6	21.8
Deeper-Cut	41.5	39.3	34.0	27.5	16.3	21.3	20.6	35.5	28.6
AOG-Simple	56.8	29.6	14.9	11.9	6.6	7.3	8.6	28.3	19.4
AOG-Seg	**58.5**	33.7	17.6	13.4	7.3	8.3	9.2	30.8	21.2
Our Model (w/o seg)	56.8	52.1	42.7	36.7	21.9	30.5	30.4	47.1	38.7
Our Model (final)	58.0	**52.1**	**43.1**	**37.2**	**22.1**	**30.8**	**31.1**	**47.6**	**39.2**

Table 1: Mean Average Precision (mAP) of Human Pose Estimation on PASCAL-Person-Part.

reported. Each groundtruth can only be matched to one estimated pose configuration. Unassigned pose configurations in C_I^{pose} are all treated as false positives.

We compare our method with two other state-of-the-art approaches: (1) Chen & Yuille [7], a tree-structured model designed specifically for single-person estimation in presence of occlusion, using unary scores and image-dependent pairwise terms based on DCNN features; (2) Deeper-Cut [15], an integer linear programming model that jointly performs multi-person detection and multi-person pose estimation. These two methods both use strong graphical assembling models. We also build two other baselines, which use simple And-Or graphs for assembling instead of the FCRF in our model. One is "AOG-Simple", which only uses geometric connectivity between neighbouring joints. The other one is "AOG-Seg", which adds part segment consistency features to "AOG-Simple". The part segment consistency features are the same as the segment-joint smoothness feature we use in the FCRF. To test the effectiveness of our proposed part segment consistency, we also list the result of our model w/o the consistency features ("Our Model (w/o seg)"). The results are shown in Tab. 1. Our model outperforms all the other methods, and by comparing our model with "AOG-Simple" and "AOG-Seg", we can see that a good assembling model is really necessary for challenging multi-person images like those in PASCAL.

Our proposed part segment consistency features not only help the overall pose estimation results, but also improve the accuracy of the detailed joint localization. Previous evaluation metrics (*e.g.* PCP, PCK and mAP) treat any joint estimate within a certain distance of the groundtruth to be correct, and thus they do not encourage joint estimates to be as close as possible to the groundtruth. Therefore, we design a new evaluation metric called Average Distance of Keypoints (ADK). For each groundtruth pose configuration, we compute its reference scale to be half of the distance between the forehead and neck, then find the only pose configuration estimate among the generated pose configuration proposals that has the highest overlap with the groundtruth configuration. For each joint that is visible in both the groundtruth configuration and the estimated configuration, the relative distance (w.r.t. the reference scale) between the estimated location and the groundtruth location is computed. Finally, we compute the average distance for each joint type across all the testing images.

Method	Forehead	Neck	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Ave.
Chen & Yuille	37.5	29.7	51.6	65.9	72.0	70.5	79.9	78.6	60.7
Deeper-Cut	32.1	30.9	37.5	44.6	53.5	53.9	65.8	67.8	48.3
AOG-Simple	33.0	33.2	66.7	82.3	90.5	89.7	101.3	101.1	74.7
AOG-Seg	32.2	31.6	59.8	72.4	85.1	85.7	97.1	92.7	69.6
Our Model (w/o seg)	27.7	26.9	33.1	40.2	47.3	51.8	54.6	53.4	41.9
Our Model (final)	**26.9**	**26.1**	**32.7**	**39.5**	**45.3**	**50.9**	**52.3**	**51.8**	**40.7**

Table 2: Average Distance of Keypoints (ADK) (%) of Human Pose Estimation on PASCAL-Person-Part.

Method	Head	Torso	U-arms	L-arms	U-legs	L-legs	Background	Ave.
Attention [5]	81.47	59.06	44.15	42.50	38.28	35.62	93.65	56.39
HAZN [33]	80.76	60.50	45.65	43.11	41.21	37.74	93.78	57.54
Our model (VGG-16, w/o pose)	79.83	59.72	43.84	40.84	40.49	37.23	93.55	56.50
Our model (VGG-16, final)	80.21	61.36	47.53	43.94	41.77	38.00	93.64	58.06
Our model (ResNet-101, w/o pose)	84.95	67.21	52.81	51.37	46.27	41.03	94.96	62.66
Our model (ResNet-101, final)	**85.50**	**67.87**	**54.72**	**54.30**	**48.25**	**44.76**	**95.32**	**64.39**

Table 3: Mean Pixel IOU (mIOU) (%) of Human Semantic Part Segmentation on PASCAL-Person-Part.

The result is shown in Tab. 2. It can be seen that our model reduces the average distance of keypoints significantly for wrists and lower-body joints by employing consistency with semantic part segmentation.

4.3. Human Semantic Part Segmentation

We evaluate the part segmentation results in terms of mean pixel IOU (mIOU) following previous works [3, 33]. In Tab. 3, we compare our model with two other state-of-the-art methods [5, 33] as well as one inferior baseline of our own model (*i.e.* the output part label map $\mathbf{L}_s$ of the first-stage part FCN, without the help of pose information).

We also list the numbers of our model using the more advanced network architecture ResNet-101 [4] instead of VGG-16 [3] for Part FCN. It can be seen that our model surpasses previous methods and the added pose information is effective for improving the segmentation results. When using ResNet-101, our model further boosts the performance to **64.39**%.

Besides, we evaluate part segmentation w.r.t. different sizes of human instances in Tab. 4, following [33]. Our model performs especially well for small-scale people, surpassing other state-of-the-arts by over **5**%.

Method	Size XS	Size S	Size M	Size L
Attention [5]	37.6	49.8	55.1	55.5
HAZN [33]	47.1	55.3	56.8	56.0
Our model (ResNet-101, w/o pose)	40.4	54.4	60.5	62.1
Our model (ResNet-101, final)	**53.4**	**60.9**	**63.0**	**62.8**

Table 4: Mean Pixel IOU (mIOU) (%) of Human Semantic Part Segmentation w.r.t. Size of Human Instance on PASCAL-Person-Part.

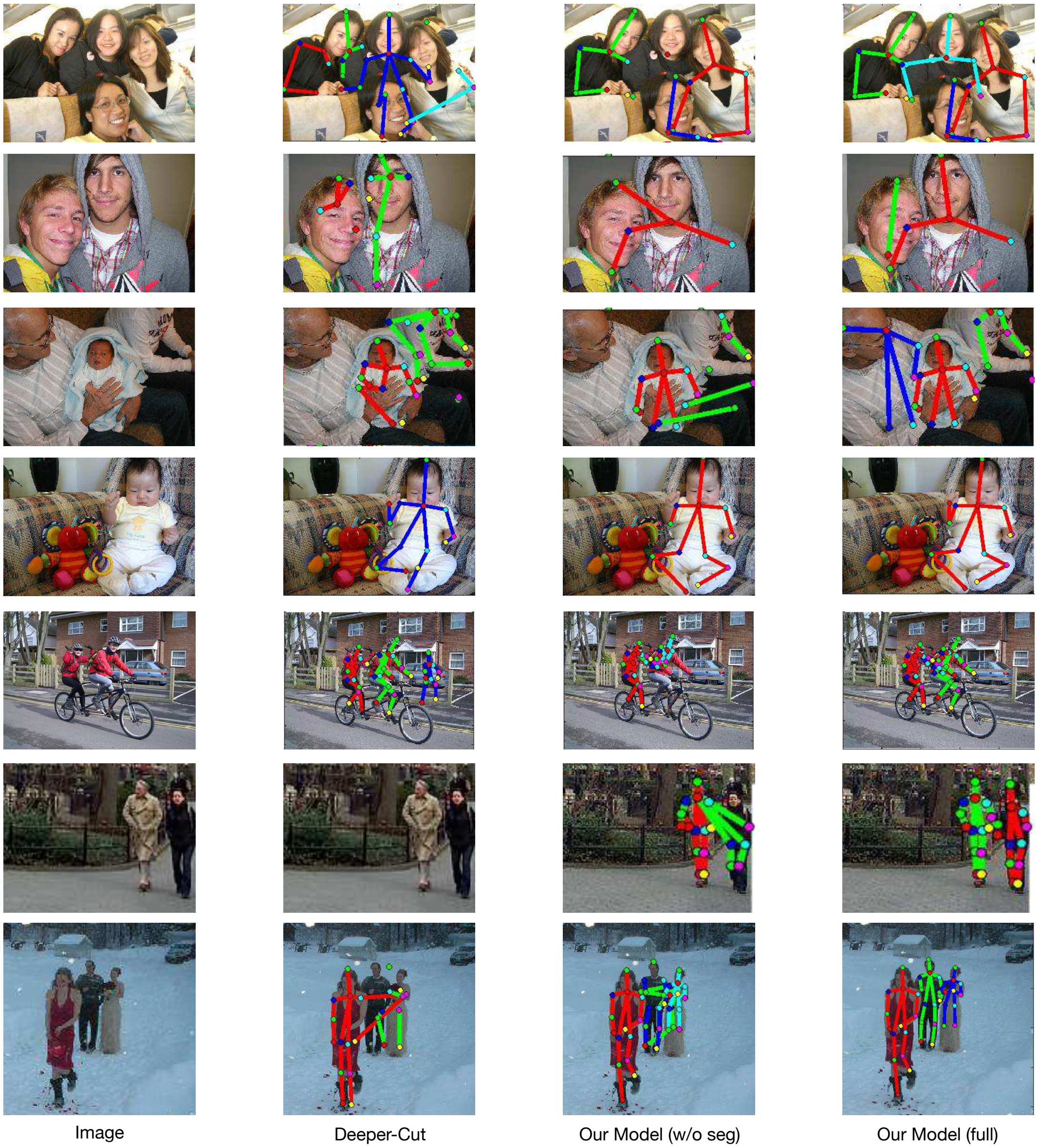

Figure 3: Visual comparison of human pose estimation on PASCAL-Person-Part [6]. Our full model is compared against Deeper-Cut [15] and a variant of our model ("Our Model (w/o seg)") that doesn't consider part segment consistency.

4.4. Qualitative Evaluation

Human pose estimation. In Fig. 3, we visually demonstrate our pose estimation results on PASCAL-Person-Part, comparing them with the recent state-of-the-art Deeper-Cut [15] and also a sub-model of ours ("Our Model (w/o seg)") which does not consider part segment consistency. This shows that our model gives more accurate prediction of heads, arms and legs, and is especially better at handling people of small scale (see the 6_{th} and 7_{th} row of Fig. 3) and extra large scale (see the first two rows of Fig. 3).

Human semantic part segmentation. Fig. 4 visually illustrates the advantages of our model over two other recent methods, Attention [5] and HAZN [33], which adopt the same basic network structure as ours. Our model estimates the overall part configuration more accurately. For example, in the 2_{rd} row of Fig. 4, we correctly labels the right arm of the person while the other two baseline methods label it as upper-leg and lower-leg. Furthermore, our model gives clearer details of arms and legs (see the last three rows of Fig. 4), especially for small-scale people.

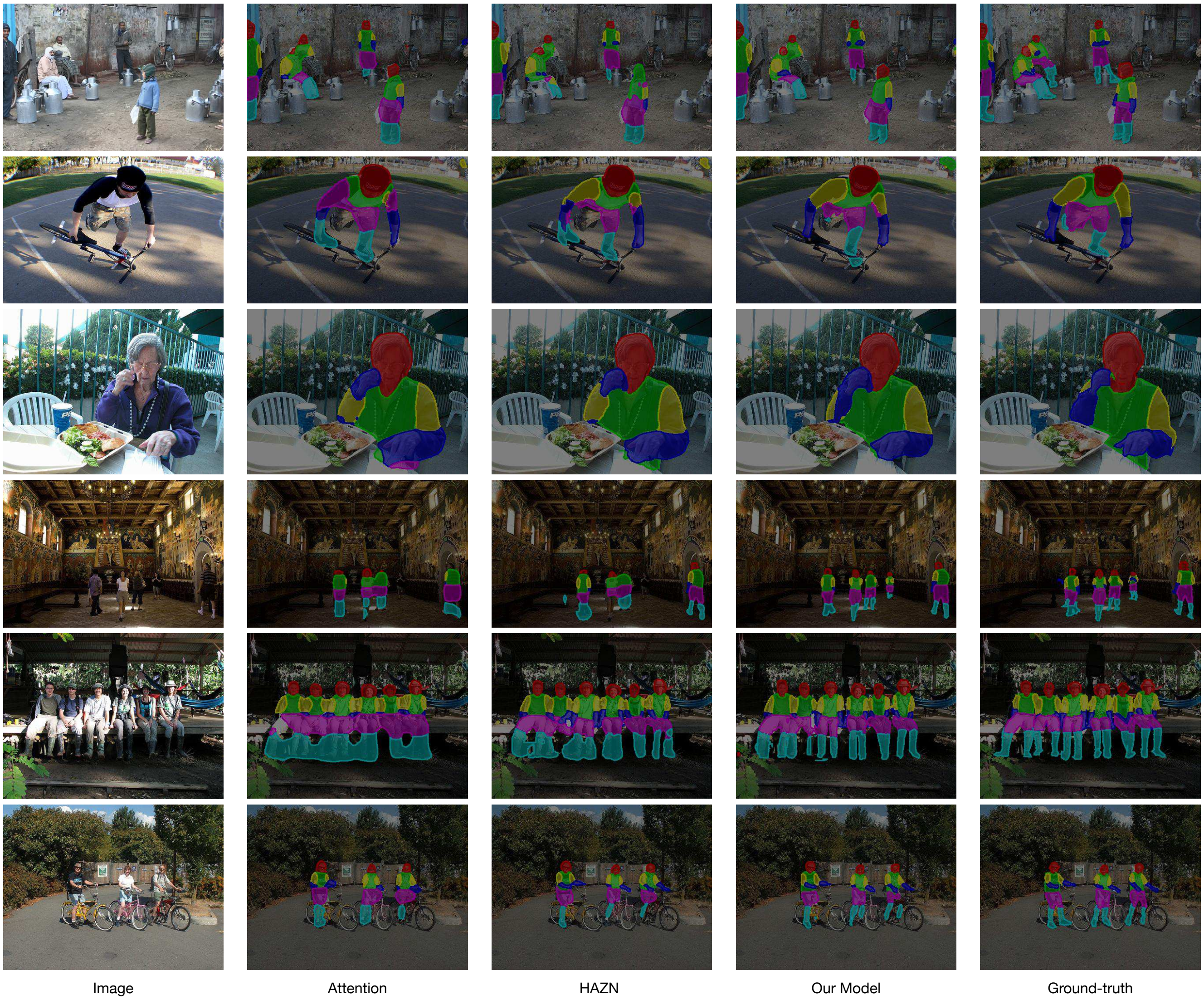

Figure 4: Visual comparison of human semantic part segmentation on PASCAL-Person-Part [6]. Our method is compared against two recent state-of-the-art methods: Attention [5] and HAZN [33].

5. Conclusion

In this paper, we demonstrate the complementary properties of human pose estimation and semantic part segmentation in complex multi-person images. We present an efficient framework that performs the two tasks iteratively and improves the results of each task. For human pose estimation, we adopt a fully-connected CRF that jointly performs human instance clustering and joint labeling, using deep-learned features and part segment based consistency features. This model gives better localization of joints, especially for arms and legs. For human semantic segmentation, we train a FCN that uses estimated pose configurations as shape and location priors, successfully correcting local confusions of people and giving clearer details of arms and legs.

We also adopt an effective "auto-zoom" strategy that deals with object scale variation for both tasks and helps reduces the inference time of the CRF by a factor of 40. We test our approach on the challenging PASCAL-Person-Part dataset and show that it outperforms state-of-the-art methods for both tasks.

6. Acknowledgements

We are deeply grateful for the support from ONR N00014-15-1-2356, NSF award CCF-1317376 and Army Research Office ARO 62250-CS, and also for the free GPUs provided by NVIDIA.

References

[1] S. Branson, G. Van Horn, S. Belongie, and P. Perona. Bird species categorization using pose normalized deep convolutional nets. *arXiv preprint arXiv:1406.2952*, 2014. 1

[2] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human pose estimation with iterative error feedback. *arXiv preprint arXiv:1507.06550*, 2015. 2

[3] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*, 2015. 1, 2, 4, 6

[4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv preprint arXiv:1606.00915*, 2016. 3, 6

[5] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille. Attention to scale: Scale-aware semantic image segmentation. *arXiv:1511.03339*, 2015. 3, 6, 7, 8

[6] X. Chen, R. Mottaghi, X. Liu, S. Fidler, R. Urtasun, and A. L. Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*, 2014. 1, 2, 5, 7, 8

[7] X. Chen and A. Yuille. Parsing occluded people by flexible compositions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 6

[8] X. Chen and A. L. Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *Advances in Neural Information Processing Systems*, pages 1736–1744, 2014. 1, 2

[9] X. Chu, W. Ouyang, H. Li, and X. Wang. Structured feature learning for pose estimation. *arXiv preprint arXiv:1603.09065*, 2016. 2

[10] J. Dong, Q. Chen, X. Shen, J. Yang, and S. Yan. Towards unified human parsing and pose estimation. In *CVPR*, 2014. 3

[11] J. Dong, Q. Chen, W. Xia, Z. Huang, and S. Yan. A deformable mixture parsing model with parselets. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3408–3415, 2013. 2

[12] M. Everingham, S. A. Eslami, L. V. Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111(1):98–136, 2014. 1

[13] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005. 2

[14] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 4

[15] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. *arXiv preprint arXiv:1605.03170*, 2016. 2, 4, 5, 6, 7

[16] S. Jones and L. Shao. Content-based retrieval of human actions from realistic video databases. *Information Sciences*, 236:56–65, 2013. 1

[17] J. Krause, B. Sapp, A. Howard, H. Zhou, A. Toshev, T. Duerig, J. Philbin, and L. Fei-Fei. The unreasonable effectiveness of noisy data for fine-grained recognition. *arXiv preprint arXiv:1511.06789*, 2015. 1

[18] L. Ladicky, P. H. Torr, and A. Zisserman. Human pose estimation using a joint pixel-wise and part-wise formulation. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3578–3585, 2013. 3

[19] Y. LeCun, K. Kavukcuoglu, C. Farabet, et al. Convolutional networks and applications in vision. In *ISCAS*, pages 253–256, 2010. 1

[20] X. Liang, X. Shen, J. Feng, L. Lin, and S. Yan. Semantic object parsing with graph lstm. *arXiv preprint arXiv:1603.07063*, 2016. 1, 3

[21] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1

[22] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 2

[23] Y. Lu, K. Boukharouba, J. Boonært, A. Fleury, and S. Lecoeuche. Application of an incremental svm algorithm for on-line human recognition from video surveillance using texture and color features. *Neurocomputing*, 126:132–140, 2014. 1

[24] L. Ma, X. Yang, Y. Xu, and J. Zhu. Human identification using body prior and generalized emd. In *2011 18th IEEE International Conference on Image Processing*, pages 1441–1444. IEEE, 2011. 1

[25] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. *arXiv preprint arXiv:1603.06937*, 2016. 1

[26] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv:1506.01497*, 2015. 2, 3

[27] L. Sigal and M. J. Black. Predicting 3d people from 2d pictures. In *International Conference on Articulated Motion and Deformable Objects*, pages 185–195. Springer, 2006. 5

[28] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler. Efficient object localization using convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 648–656, 2015. 2

[29] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, 2014. 2

[30] C. Wang, Y. Wang, and A. L. Yuille. An approach to pose-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 915–922, 2013. 1

[31] P. Wang, X. Shen, Z. Lin, S. Cohen, B. Price, and A. L. Yuille. Joint object and part segmentation using deep learned potentials. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1573–1581, 2015. 1, 2

[32] Y. Wang, D. Tran, Z. Liao, and D. Forsyth. Discriminative hierarchical part-based models for human parsing and action recognition. *Journal of Machine Learning Research*, 13(Oct):3075–3102, 2012. 1

[33] F. Xia, P. Wang, L.-C. Chen, and A. L. Yuille. Zoom better to see clearer: Human and object parsing with hierarchical auto-zoom net. In *ECCV*, 2016. 1, 2, 3, 4, 6, 7, 8

[34] F. Xia, J. Zhu, P. Wang, and A. Yuille. Pose-guided human parsing by an and/or graph using pose-context features. In *AAAI Conference on Artificial Intelligence*, 2016. 1, 2

[35] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. Parsing clothing in fashion photographs. In *CVPR*, 2012. 2, 3

[36] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg. Retrieving similar styles to parse clothing. *IEEE transactions on pattern analysis and machine intelligence*, 37(5):1028–1040, 2015. 1

[37] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011. 2

[38] N. Zhang, J. Donahue, R. Girshick, and T. Darrell. Part-based r-cnns for fine-grained category detection. In *European Conference on Computer Vision*, pages 834–849. Springer, 2014. 1