Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Pose-Guided Human Parsing by an AND/OR Graph Using Pose-Context Features

Fangting Xia and **Jun Zhu** and **Peng Wang*** and **Alan L. Yuille**
Department of Statistics
University of California, Los Angeles
Los Angeles, California 90095

Abstract

Parsing human into semantic parts is crucial to human-centric analysis. In this paper, we propose a human parsing pipeline that uses pose cues, i.e., estimates of human joint locations, to provide pose-guided segment proposals for semantic parts. These segment proposals are ranked using standard appearance cues, deep-learned semantic feature, and a novel pose feature called pose-context. Then these proposals are selected and assembled using an And-Or graph to output a parse of the person. The And-Or graph is able to deal with large human appearance variability due to pose, choice of clothes, etc. We evaluate our approach on the popular Penn-Fudan pedestrian parsing dataset, showing that it significantly outperforms the state-of-the-arts, and perform diagnostics to demonstrate the effectiveness of different stages of our pipeline.

Introduction

The goal of human parsing is to partition the human body into different semantic parts such as hair, face, torso, arms, and legs. This provides rich descriptions for human-centric image analysis which is increasingly important for many computer vision applications such as content-based image/video retrieval (Weber et al. 2011), person re-identification (Ma et al. 2011; Cheng et al. 2011), video surveillance (Yang and Yu 2011), action recognition (Wang et al. 2012; Zhu et al. 2013; Wang, Wang, and Yuille 2013), and clothes fashion recognition (Yamaguchi et al. 2012). But human parsing is very challenging in real-life scenarios due to extreme variability in human appearance and shape caused by human poses, clothes types, and occlusion/self-occlusion patterns.

The leading approach to human parsing uses a segment-based graphical model, which first generates segment/region proposals for human parts (Yamaguchi et al. 2012; Yang, Luo, and Lin 2014) based on low-level appearance cues (e.g., similarity of color/texture, or grouping edges) and then selects and integrates these proposed segments using a graphical model (Bo and Fowlkes 2011; Yang, Luo, and Lin 2014; Dong et al. 2014; Liu et al. 2015). But the low-level cue for part proposal generation and ranking is problematic for complex images. To select proposals, a proposed solution is to use pose information. But this pose information is only used at the final stage (Dong et al. 2014) and so cannot correct errors made in the original proposal generation. As illustrated in Fig. 1 and Fig. 2, our approach uses *pose-guided-proposals* and *pose-context*, obtained from high-level pose analysis, to improve the quality of part proposals (in Fig.1(b)), to provide effective features for ranking proposals (in Fig.1(c)), and to enable us to select and integrate these proposals using a graphical model (in Fig.1(d)).

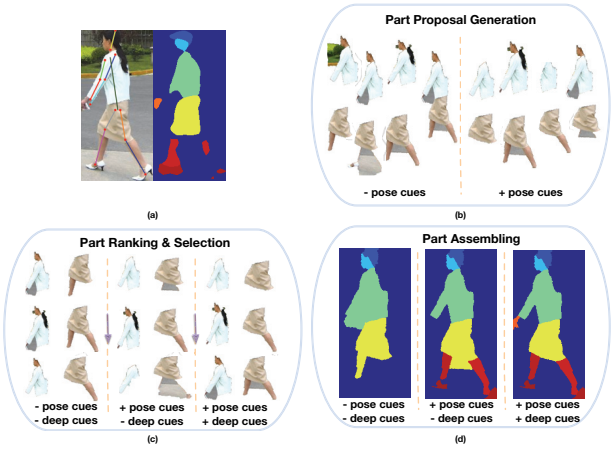


Figure 1: Human parsing using pose (pose-guided-proposals and pose-context features) and deep-learned part semantic cues. (a) Left: original image with estimated pose joints. Right: the inferred part label map by deep semantic cues. (b) Pose-guided part proposal generation. Left: without pose information. Right: with pose information. (c) Top-ranked part segments after part ranking and selection. Left: without pose-context and deep semantic features. Middle: using pose-context feature only. Right: using both pose-context and deep semantic features. (d) Final parsing results. Left: without pose-context and deep semantic features. Middle: using pose-context feature only. Right: using both pose-context and deep semantic features.

The overall strategy of our approach is illustrated in Fig. 2. Given an input image, we first (bottom left) use a

*equal contribution with Jun Zhu.

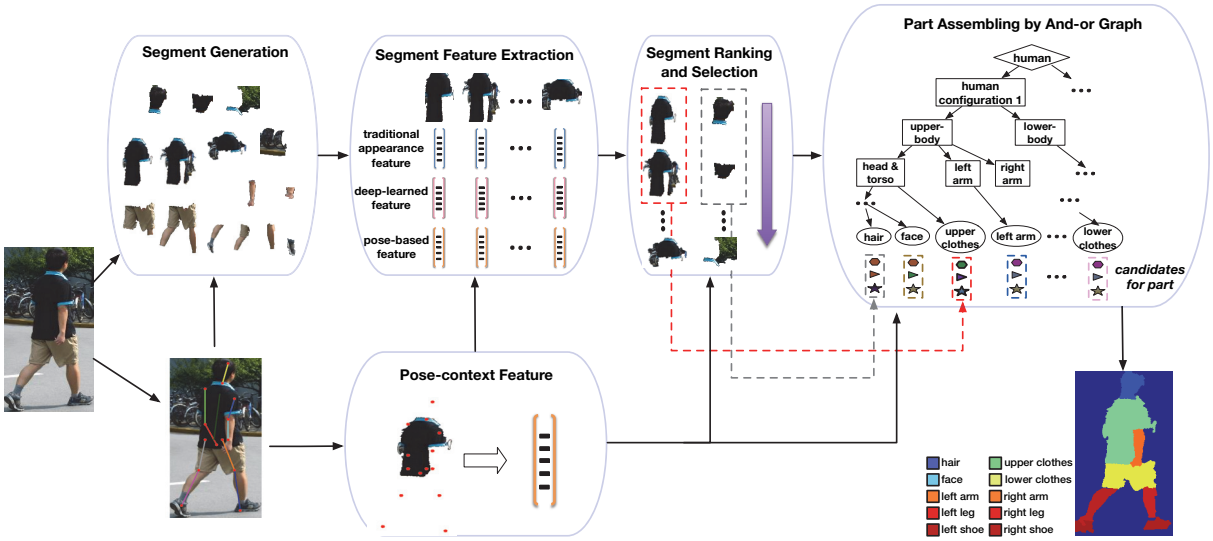


Figure 2: Illustration of our human parsing pipeline.

state-of-the-art pose estimation algorithm (Chen and Yuille 2014) to estimate the locations of the joints and other salient parts of humans. We use the estimates of the joint positions to obtain *pose-guided-proposals* for part segments (top left) based on the intuition that part segments should be correlated to joint positions (e.g., the lower-arm should appear between the wrist and the elbow), which yields a limited set of proposals with high recall. Next we compute rich feature descriptors for each segment proposal, including a novel *pose-context* feature which captures spatial/geometrical relationship between a proposed segment and the estimated human pose joints. We also use standard appearance features and complementary deep-learned part semantic features computed by a fully convolutional network (FCN) (Long, Shelhamer, and Darrell 2015; Hariharan et al. 2015; Chen et al. 2015; Tsogkas et al. 2015). Then we rank the segment proposals based on these features and select the top-ranked ones. This leaves a small number of high-quality proposals for each part category which are used as input to the part assembling stage.

For part assembling, we propose an And-Or graph (AOG) (Zhu and Mumford 2007; Zhu et al. 2008; 2012; Wang and Yuille 2015), which is an efficient way to represent the large variability of human appearances. We perform inference over this AOG to select and combine part segment proposals so as to parse the human body. Compared with traditional AOGs, our AOG has more flexible and efficient structure (i.e. each leaf node allows arbitrary number of data-mined part subtypes) and includes an extension of the pose-context feature as a pairwise term to measure the compatibility of adjacent parts. Unlike the local pose features in Dong's AOG (2014), our AOG measures the consistency between pose and segment both locally and globally.

We evaluate our method on a popular pedestrian parsing benchmark dataset (i.e., *Penn-Fudan* (Wang et al. 2007)), and show that our approach outperforms other state-of-the-arts by a significant margin.

Pose-Guided Human Parsing Pipeline

Given a pedestrian image I, we first adopt a state-of-the-art pose estimation approach (Chen and Yuille 2014) to estimate human pose joints $\mathcal{L} = \{l_1, l_2, \cdots, l_{14}\}$, where l_j denotes the location of the j-th pose joint. Here we use the same 14 joints as those commonly used in the human pose estimation literature (Yang and Ramanan. 2011; Chen and Yuille 2014). As shown in Fig. 2, based on the human pose cues, our human parsing pipeline has three successive steps: *part segment proposal generation*, *part proposal selection*, and *part assembling*. We will introduce the first two steps below, and elaborate on our AOG-based part assembling method in the next section.

Pose-guided part segment proposal generation. To generate part segment proposals, we modify the RIGOR algorithm (Humayun, Li, and Rehg 2014), which can efficiently generate segments aligning with object boundaries given user defined initial seeds and cutting thresholds. In this paper, we propose to generate the seeds based on the estimated pose joint locations. Specifically, given the observation that part segments tend to be surrounding corresponding pose joints, for each joint we sample a set of seeds at the 5×5 grid locations over a 40×40 image patch centered at this joint. We use 8 different cutting thresholds, yielding about 200 segment proposals for each joint. Combining proposals from all the joints, we further prune out duplicate segments (with intersect-over-union (IOU) ≥ 0.95 as threshold) and construct a segment pool $\mathcal{S} = \{s_1, s_2, \cdots, s_N\}$ that contains around 800 segment proposals for each image. We use these segments as candidate part segments in the latter two steps.

Part proposal selection. We consider the following image features for each segment proposal $s_i \in \mathcal{S}$: (i) $\phi^{o2p}(s_i)$,

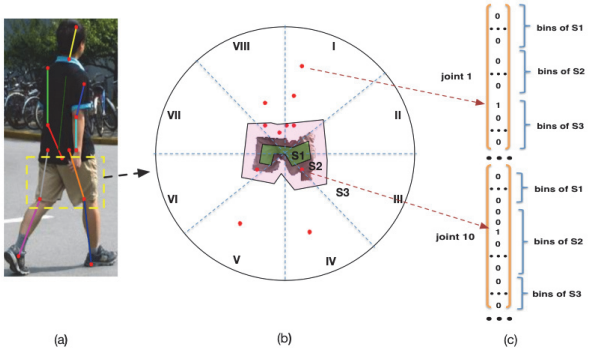


Figure 3: Illustration of the proposed pose-context feature.

Figure 4: The learned prototypes/clusters for part category *face*. We show exemplar images for 3 out of 6 clusters. Cluster (1): frontal face or back face. Cluster (2): frontal/back face on the left. Cluster (3): side face on the left. The other clusters correspond to the symmetric patterns w.r.t. those shown here.

a second order pooling (O2P) feature (Carreira et al. 2012) for describing appearance cues; (ii) $\phi^{skin}(s_i)$, an appearance feature (Khan et al. 2010) capturing skin color cues; (iii) $\phi^{pose}(s_i, \mathcal{L})$, a pose-context feature we propose in this paper, which measures the spatial relationship between the segment s_i and the predicted pose joint configuration $\mathcal{L}$; (iv) $\phi^{c-pose}(s_i, \mathcal{L})$, a non-linearly coded version of $\phi^{pose}(s_i, \mathcal{L})$; (v) $\phi^{fcn}(s_i, \mathcal{H})$, a deep-learned semantic feature using FCN, which measures the compatibilities between the segment image patch and high-level part semantic cues from FCN.

We now describe the proposed pose-context feature $\phi^{pose}(s_i, \mathcal{L})$. As shown in Fig. 3, centered at s_i, the image is equally divided into eight orientations (I – VIII) and three region scales (S1, S2 and S3), yielding 24 spatial bins in total. Then each joint $l_j \in \mathcal{L}$ falls into one of these spatial bins, producing a binary feature to quantize the spatial relationship of l_j w.r.t. s_i. After that, we concatenate binary features of all the joints, and obtain a $24 \times 14 = 336$ dimensional pose-context feature to describe the spatial relationship of s_i w.r.t. $\mathcal{L}$. Specifically, S1 and S2 are the regions eroded and dilated by 10 pixels from the segment's boundary respectively. S3 is the rest region of image. This segment-dependent definition of region scales depicts semantically meaningful geometric cues from the predicted pose information, e.g. the lower boundary of the short skirt segment should be around the knee joints. The three-scale design (rather than using the segment edge alone) makes the feature robust to pose estimation errors.

The pose-context feature can be highly non-linear in the feature space, which might be suboptimal for linear classifiers/regressors. This motivates us to apply non-linear coding technology (Yang et al. 2009; Wang et al. 2010) on the pose-context feature to achieve linearity. We adopt a soft-assignment quantization (SAQ) coding method (Liu, Wang, and Liu 2011) to encode the pose-context feature into its coded version $\phi^{c-pose}(s_i, \mathcal{L})$, with a dictionary of pose-guided part prototypes $\mathcal{D} = \{\mathbf{b}_m\}_{m=1}^{N_\mathcal{D}}$, learned via K-means clustering algorithm on the pose-context feature representation of ground-truth part segment examples. Specifically, to balance K different part categories, we separately perform clustering and obtain $N_p = 6$ prototypes/clusters for each part category, resulting in a dictionary of $N_\mathcal{D} = K \times N_p$ codewords. Given $\mathcal{D}$, we compute the Euclidean distance between original pose-context feature of s_i and each prototype $\mathbf{b}_m$: $d_{i,m} = \| \phi^{pose}(s_i, \mathcal{L}) - \mathbf{b}_m \|$. Thus $\phi^{c-pose}(s_i, \mathcal{L})$ is formally defined as the concatenation of both the normalized and un-normalized codes w.r.t. $\mathcal{D}$:

$$\phi^{c-pose}(s_i, \mathcal{L} \,|\, \mathcal{D}) = [a_{i,1}, \cdots, a_{i,N_\mathcal{D}}, a'_{i,1}, \cdots, a'_{i,N_\mathcal{D}}]^{\mathrm{T}}, \quad (1)$$

where $a_{i,m} = \exp^{(-\lambda d_{i,m})}$ and $a'_{i,m} = \frac{a_{i,m}}{\sum_{j=1}^{N_\mathcal{D}} a_{i,j}}$ denote the un-normalized and normalized code values w.r.t. $\mathbf{b}_m$ respectively. λ is a hyper-parameter of our coding method. The coded pose-context feature is adopted in training the SVR models for part proposal selection. The learned part prototypes, which generally correspond to different viewpoints of a part or different appearance patterns of a part (e.g. long pants or skirts for the lower-clothes category), are used to define part subtypes in our AOG. As illustrated in Fig. 4, the learned face prototypes generally correspond to different typical views of the face category. Besides, we propose to encode the pairwise pose-context feature (i.e. concatenated pose-context features of a pair of candidate segments), used as a pairwise term in our AOG design. We perform clustering separately for each adjacent part pair and learn a class-specific dictionary for this pairwise pose-context feature. In this paper, the dictionary size is set by $N_{pp} = 8$ for each part pair. As visualized in Fig. 5, the learned part-pair prototypes are very meaningful which capture typical viewpoints and part type co-occurrence patterns for adjacent parts.

For the deep-learned semantic feature, we train a FCN-16s deep network (Hariharan et al. 2015; Wang et al. 2015) with the output to be the part ground truth map, and then produce pixel-wise part potential maps $\mathcal{H}$, from which binary part label masks $\mathcal{B}$ can be obtained via argmax over the potential maps. Thus, for a segment s_i, this deep feature

Figure 5: The learned prototypes/clusters for the adjacent part pair *upper-clothes and lower-clothes*. We show 3 out of 8 clusters. Cluster (1): the person with short sleeved upper-clothes and short pants. Cluster (2): the person with short sleeved upper-clothes and long pants. Cluster (3): the person with long sleeved upper-clothes and long pants.

$\phi^{fcn}(s_i, \mathcal{H})$ consists of three components: (1) the mean value inside s_i of $\mathcal{H}$ for each part class; (2) the mean value along the contour of s_i from $\mathcal{H}$ for each part class; (3) The IoU value between s_i and $\mathcal{B}$ for each part class.

Our final feature descriptor of s_i is the concatenation of the aforementioned features, i.e.,

$$\begin{aligned}\phi(s_i, \mathcal{L}, \mathcal{H}) = [&\phi^{o2p}(s_i),\ \phi^{skin}(s_i),\ \phi^{fcn}(s_i, \mathcal{H}),\\ &\phi^{pose}(s_i, \mathcal{L}),\ \phi^{c-pose}(s_i, \mathcal{L})]^{\mathrm{T}}\end{aligned} \tag{2}$$

On basis of this hybrid feature representation, we train a linear support vector regressor (SVR) (Carreira et al. 2012) for each part category. Let P denote the total number of part categories and $p \in \{1, 2, \cdots, P\}$ denote the index of a part category. The target variable for training SVR is the IoU value between the segment proposal and ground-truth label map of part p. The output of SVR is given by Equ. (3).

$$g^p(s_i|\mathcal{L}, \mathcal{H}) = {\beta_p}^{\mathrm{T}} \phi(s_i, \mathcal{L}, \mathcal{H}), \tag{3}$$

where β_p is the model parameter of SVR for the p-th part category. Thus, for any part p, we rank the segment proposals in $\mathcal{S}$ based on their SVR scores $\{g^p(s_i) \,|\, s_i \in \mathcal{S}\}$. Finally, we select the top-n_p scored segments separately for each part category and combine the selected segment proposals from all part categories to form a new segment pool $\tilde{\mathcal{S}} \subseteq \mathcal{S}$.

Part Assembling with And-Or Graph

There are two different groups of classes (i.e., *parts* and *part compositions*) in our AOG model: the part classes are the finest-level constituents of human body; the part compositions correspond to intermediate concepts in the hierarchy of semantic human body constituents. Specifically, we define them as follows. *Parts*: hair, face, full-body clothes, upper-clothes, left/right arm, lower-clothes, left/right leg skin, left/right shoe. *Part Compositions*: head, head & torso, upper-body, left/right leg, human body.

To assemble the selected part segments, we develop a compositional AOG model as illustrated in Fig. 6, which facilitates flexible composition structure and standard learning/inference routines. Let P and C denote the number of parts and the number of part compositions respectively. Formally, our AOG model is defined as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \mathcal{T} \cup \mathcal{N}$ denotes a set of vertices and $\mathcal{E}$ refers to the set of edges associated. Meanwhile, $\mathcal{T} = \{1, 2, \cdots, P\}$ and $\mathcal{N} = \{P+1, P+2, \cdots, P+C\}$ denote the set of part indices and the set of part composition indices respectively. In our AOG, each leaf vertex $p \in \mathcal{T}$ represents one human body part and each non-leaf vertex $c \in \mathcal{N}$ represents one part composition. The root vertex corresponds to the whole human body while the vertices below correspond to the part compositions or parts at various semantic levels. Our goal is to parse the human body into a series of part compositions and parts, which is in a hierarchical graph instantiated from the AOG model.

The vertex of our AOG is a nested subgraph as illustrated at the bottom of Fig. 6. For a leaf vertex $p \in \mathcal{T}$, it includes one Or-node followed by a set of terminal nodes as its children. The terminal nodes correspond to different part subtypes learned by clustering the pose-context feature of training part segments (see last section for details), and the Or-node represents a mixture model indicating the selection of one part subtype from terminal nodes. Formally, we define a state variable $z_p \in \{0, 1, 2, \cdots, K_p\}$ to indicate that the Or-node selects the z_p-th terminal node as the part subtype for leaf vertex p. As an example of a green node in Fig. 6, the lower-clothes part can select one kind of subtype (e.g. long pants or skirt) from its candidate part subtypes. In addition, there is one special terminal node representing the invisibility of the part due to occlusion/self-occlusion, which corresponds to the state $z_p = 0$. For a non-leaf vertex $c \in \mathcal{N}$, it includes one Or-node linked by a set of And-nodes plus one terminal node. The Or-node of non-leaf vertex represents this part composition has several different ways of decompositions into smaller parts and/or part compositions. The And-node corresponds to one possible configuration of the decomposition of c. As shown in Fig. 6, the non-leaf vertex head can be composed by one of several different configurations of two child vertices (i.e., face and hair). Similar to the leaf vertices, we also induce a state variable $z_c \in \{0, 1, 2, \cdots, K_c\}$ to indicate that the Or-node of part composition c selects the z_c-th And-node as the configuration of child vertices for $z_c \neq 0$ or this part composition is invisible when $z_c = 0$.

Further, we define another state variable y to indicate the selection of segment from the candidate pool of a part or part composition. For a leaf vertex $p \in \mathcal{T}$, $y_p \in \{0, 1, 2, \cdots, n_{p,z_p}\}$ represents that the part p selects the y_p-th segment proposal (i.e., $s_{y_p}^{p,z_p}$) from the segment pool $\tilde{\mathcal{S}}_{p,z_p}$, outputted by its segment ranking model on subtype z_p. Meanwhile, $y_p = 0$ is a special state which coincides with the invisibility pattern

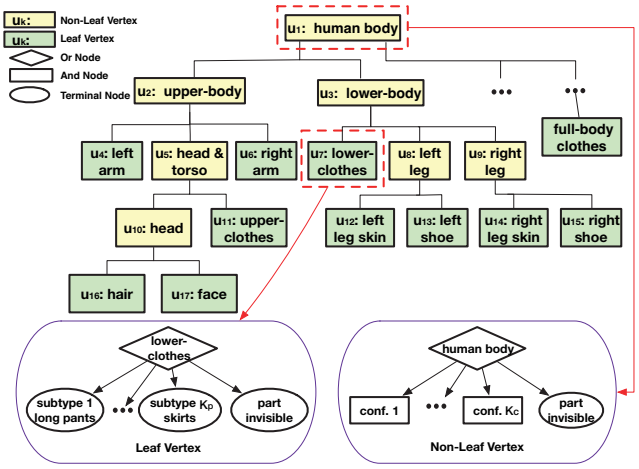


Figure 6: Illustration of the architecture of our AOG model.

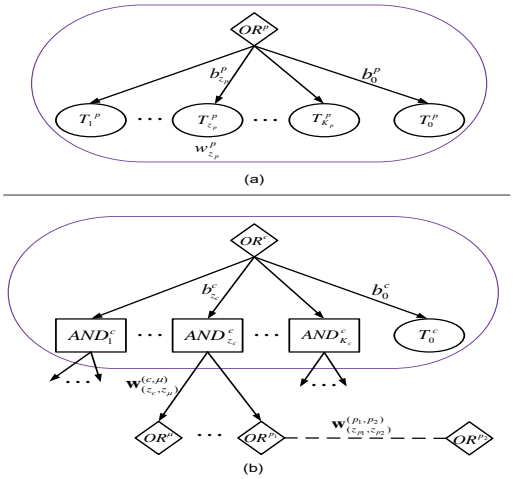


Figure 7: Illustration on the structure of vertices in AOG. (a) leaf vertex; (b) non-leaf vertex. The symbols of *OR*, *AND* and *T* represent the Or-node, And-node and terminal node respectively. Please see Eqn. (6) and Eqn. (7) about the notations of model parameters.

of part p (i.e., $z_p = 0$). To make the notations consistent, we use s_0^{p,z_p} to represent an "null" segment for part invisibility. For a non-leaf vertex $c \in \mathcal{N}$, $y_c \in \{0, 1, 2, \cdots, n_c\}$ indicates a segment $s_{y_c}^{c,z_c} \in \tilde{\mathcal{S}}_{c,z_c}$ is selected, where $s_{y_c}^{c,z_c}$ is obtained by the union of its child vertices' candidate segments and $\tilde{\mathcal{S}}_{c,z_c}$ denotes the candidate segment pool for the z_c And-node. When $y_c = 0$, likewise, the s_0^{c,z_c} represents a null segment indicating the invisibility pattern of part composition c. Let $Ch(c, z_c)$ denote the set of child vertices for part composition c and configuration z_c. Formally, $s_{y_c}^{c,z_c}$ is defined by Equ. (4), where $\cup$ represents a pixel-wise union operation of combing the child vertices' segment masks to generate a new segment.

$$s_{y_c}^{c,z_c} = \cup_{\mu \in Ch(c,z_c)} s_{y_\mu}^{\mu,z_\mu}, \tag{4}$$

Part Composition (c)	Adjacent Part Pairs ($\mathcal{R}_c$)
human body	(upper-clothes, lower-clothes), (full-body clothes, left leg skin) (full-body clothes, right leg skin)
head	(hair, face)
head & torso	(upper-clothes, hair), (upper-clothes, face) (full-body clothes, hair), (full-body clothes, face)
upper-body	(left arm, upper-clothes), (right arm, upper-clothes) (left arm, full-body clothes), (right arm, full-body clothes)
lower-body	(lower-clothes, left leg skin), (lower-clothes, right leg skin) (lower-clothes, left shoe), (lower-clothes, right shoe)
left leg	(left leg skin, left shoe)
right leg	(right leg skin, right shoe)

Table 1: The list of adjacent part pairs.

Let $\mathbf{Y} = (y_1, y_2, \cdots, y_P, y_{P+1}, y_{P+2}, \cdots, y_{P+C})$ and $\mathbf{Z} = (z_1, z_2, \cdots, z_P, z_{P+1}, z_{P+2}, \cdots, z_{P+C})$ denote the structural solution of AOG. We define a global score function of AOG $F(\mathbf{Y}, \mathbf{Z} \,|\, \tilde{\mathcal{S}}, \mathcal{L}, \mathcal{H})$ (here $\tilde{\mathcal{S}} = \bigcup_{p \in \mathcal{T}, z_p \neq 0} \tilde{\mathcal{S}}_{p,z_p}$) to measure the compatibility between $(\mathbf{Y}, \mathbf{Z})$ and $(\tilde{\mathcal{S}}, \mathcal{L}, \mathcal{H})$ for image I, which can be calculated as shown in Equ. (5), where $f(y_p, z_p)$ is a local score function of leaf vertex p, which consists of only one unary term, while $f(y_c, z_c, \{(y_\mu, z_\mu) : \mu \in Ch(c, z_c)\})$ denotes a score function of non-leaf vertex c, which consists of two pairwise terms.

$$\begin{aligned} F(\mathbf{Y}, \mathbf{Z} | \tilde{\mathcal{S}}, \mathcal{L}, \mathcal{H}) = & \sum_{p \in \mathcal{T}} f(y_p, z_p) \\ & + \sum_{c \in \mathcal{N}} f(y_c, z_c, \{(y_\mu, z_\mu) : \mu \in Ch(c, z_c)\}) \end{aligned} \tag{5}$$

For each leaf vertex $p \in \mathcal{T}$ (i.e., a part), we compute $f(y_p, z_p)$ by Equ. (6), in which $w_{z_p}^p$ and $b_{z_p}^p$ denote the weight and bias parameters of unary term for part p respectively. Particularly, b_0^p is the bias parameter for the invisibility pattern of p. Besides, $g_{z_p}^p$ is dependent on the part subtype z_p, implying the regression models defined in Equ. (3) are trained by different parts and subtypes. Fig. 7 (a) illustrates the structure of a leaf vertex and its corresponding model parameters.

$$f(y_p, z_p) = \begin{cases} b_{z_p}^p + w_{z_p}^p \cdot g_{z_p}^p(s_{y_p}^{p,z_p} | \mathcal{L}, \mathcal{H}), & z_p \neq 0 \\ b_0^p, & z_p = 0 \end{cases} \tag{6}$$

For each non-leaf vertex $c \in \mathcal{N}$ (i.e., a part composition), we compute $f(y_c, z_c, \{(y_\mu, z_\mu) : \mu \in Ch(c, z_c)\})$ by Eqn. (7),

$$\begin{aligned} & f(y_c, z_c, \{(y_\mu, z_\mu) : \mu \in Ch(c, z_c)\}) \\ = & \begin{cases} b_{z_c}^c + u(y_c, z_c, \{(y_\mu, z_\mu) : \mu \in Ch(c, z_c)\}), & z_c \neq 0 \\ b_0^c, & z_c = 0 \end{cases} \end{aligned} \tag{7}$$

where

$$\begin{aligned} & u(y_c, z_c, \{(y_\mu, z_\mu) : \mu \in Ch(c, z_c)\}) \\ = & \sum_{\mu \in Ch(c,z_c)} \mathbf{w}_{(z_c, z_\mu)}^{(c,\mu)\,\mathrm{T}} \varphi(s_{y_c}^{c,z_c}, s_{y_\mu}^{\mu,z_\mu}) \\ & + \sum_{(p_1,p_2) \in \mathcal{R}_c} \mathbf{w}_{(z_{p_1}, z_{p_2})}^{(p_1,p_2)\,\mathrm{T}} \psi(s_{y_{p_1}}^{p_1,z_{p_1}}, s_{y_{p_2}}^{p_2,z_{p_2}} | \mathcal{L}). \end{aligned} \tag{8}$$

Concretely, Eqn. (7) can be divided into three terms:

(1) the bias term of selecting z_c for the Or-node, i.e. $b^c_{z_c}$. b^c_0 is the bias parameter when part composition c is invisible (In this case, all the descendant vertices are also invisible and thus the latter two terms are zero).

(2) the sum of parent-child pairwise terms (i.e., *vertical edges*) for measuring the spatial compatibility between the segment of part composition c and the segments of its child vertices, i.e. $\sum_{\mu \in Ch(c,z_c)} \mathbf{w}^{(c,\mu)}_{(z_c,z_\mu)}{}^{\mathrm{T}} \varphi(s^{c,z_c}_{y_c}, s^{\mu,z_\mu}_{y_\mu})$, where $\varphi(s^{c,z_c}_{y_c}, s^{\mu,z_\mu}_{y_\mu})$ denotes a spatial compatibility feature of segment pair $(s^{c,z_c}_{y_c}, s^{\mu,z_\mu}_{y_\mu})$ and $\mathbf{w}^{(c,\mu)}_{(z_c,z_\mu)}$ refers to corresponding weight vector. Specifically, φ is defined by $[dx; dx^2; dy; dy^2; ds; ds^2]$, in which dx, dy represent the spatial displacement between the center locations of two segments while ds is the scale ratio of them.

(3) the sum of pairwise terms (i.e., *side-way edges*) for measuring the geometric compatibility on all segment pairs specified by an adjacent part-pair set $\mathcal{R}_c$, which defines a couple of adjacent part pairs for c (e.g., for the part composition of lower body, we consider lower-clothes and leg skin to be an adjacent part pair). Tab. 1 lists the adjacent part pairs for each non-leaf vertex. To avoid double counting in recursive computation of Eqn. (8), $\mathcal{R}_c$ only includes the relevant part pairs which have at least one child vertex of c. This side-way pairwise potential corresponds to $\sum_{(p_1,p_2)\in\mathcal{R}_c} \mathbf{w}^{(p_1,p_2)}_{(z_{p_1},z_{p_2})}{}^{\mathrm{T}} \psi(s^{p_1,z_{p_1}}_{y_{p_1}}, s^{p_2,z_{p_2}}_{y_{p_2}}|\mathcal{L})$ in Eqn. (8), where $\psi(s^{p_1,z_{p_1}}_{y_{p_1}}, s^{p_2,z_{p_2}}_{y_{p_2}}|\mathcal{L})$ represents a geometric compatibility feature of segment pair $(s^{p_1,z_{p_1}}_{y_{p_1}}, s^{p_2,z_{p_2}}_{y_{p_2}})$ and $\mathbf{w}^{(p_1,p_2)}_{(z_{p_1},z_{p_2})}$ is corresponding weight vector. In this paper, we use a coded version of pose-context feature for ψ. Specifically, we adopt the same coding process as in $\phi^{c-pose}(s_i, \mathcal{L})$ but using the concatenated pose-context features for segment pair $(s^{p_1,z_{p_1}}_{y_{p_1}}, s^{p_2,z_{p_2}}_{y_{p_2}})$.

In Fig. 7 (b), we illustrate the structure of a non-leaf vertex and its corresponding model parameters.

Learning and Inference for AOG

The score function in Equ. (5) is a generalized linear model w.r.t. its parameters. We can concatenate all the model parameters to be a single vector $\mathbf{W}$ and rewrite Equ. (5) by $F(\mathbf{Y},\mathbf{Z}|\mathcal{L},\tilde{\mathcal{S}},\mathcal{H}) = \mathbf{W}^{\mathrm{T}}\mathbf{\Phi}(\mathcal{L},\tilde{\mathcal{S}},\mathcal{H},\mathbf{Y},\mathbf{Z})$. $\mathbf{\Phi}(\mathcal{L},\tilde{\mathcal{S}},\mathcal{H},\mathbf{Y},\mathbf{Z})$ is a re-organized sparse vector gathering all the features based on the structural state variable $(\mathbf{Y},\mathbf{Z})$. In our AOG model, $\mathbf{Z}$ determines the topological structure of a feasible solution (i.e., parse tree), and $\mathbf{Y}$ specifies the segments selected for the vertices of this parse tree. Given a set of labelled examples $\{(\mathbf{Y}_n,\mathbf{Z}_n)\,|\,n=1,2,\cdots,J\}$, we formulate a structural max-margin learning problem on $\mathbf{W}$ (Equ. (9)),

$$\min_{\mathbf{W}} \tfrac{1}{2}\mathbf{W}^{\mathrm{T}}\mathbf{W} + C\sum_{n=1}^{J}\xi_n \qquad (9)$$

$$\mathbf{W}^T\mathbf{\Phi}(\mathcal{L}_n,\tilde{\mathcal{S}}_n,\mathcal{H}_n,\mathbf{Y}_n,\mathbf{Z}_n) - \mathbf{W}^T\mathbf{\Phi}(\mathcal{L}_n,\tilde{\mathcal{S}}_n,\mathcal{H}_n,\mathbf{Y},\mathbf{Z}) \geq \Delta(\mathbf{Y}_n,\mathbf{Z}_n,\mathbf{Y},\mathbf{Z}) - \xi_n,\ \ s.t.\ \forall\ \mathbf{Y} \text{ and } \mathbf{Z},$$

where $\Delta(\mathbf{Y}_n,\mathbf{Z}_n,\mathbf{Y},\mathbf{Z})$ is a structural loss function to penalize a hypothesized parse tree $(\mathbf{Y},\mathbf{Z})$ different from ground truth annotation $(\mathbf{Y}_n,\mathbf{Z}_n)$. Similar to Yadollahpour, Batra, and Shakhnarovich (2013), we adopt a relative loss as in Equ. (10), i.e., the loss of hypothesized parse tree relative to the best one $(\mathbf{Y}^*,\mathbf{Z}^*)$ that could be found from the candidate pool. That is,

$$\Delta(\mathbf{Y}_n,\mathbf{Z}_n,\mathbf{Y},\mathbf{Z}) = \delta(\mathbf{Y}_n,\mathbf{Z}_n,\mathbf{Y},\mathbf{Z}) - \delta(\mathbf{Y}_n,\mathbf{Z}_n,\mathbf{Y}^*,\mathbf{Z}^*), \qquad (10)$$

where $\delta(\mathbf{Y},\mathbf{Z},\mathbf{Y}',\mathbf{Z}') = \sum_{p\in\mathcal{T}} IoU(s^{p,z_p}_{y_p}, s^{p,z'_p}_{y'_p})$ is a function of measuring the part segmentation difference between any two parse trees $(\mathbf{Y},\mathbf{Z})$ and $(\mathbf{Y}',\mathbf{Z}')$. In this paper, we employ the commonly-used cutting plane algorithm (Joachims, Finley, and Yu 2009) to solve this structural max-margin optimization problem of Equ. (9).

For inference on AOG models, dynamic programming (DP) is commonly used in the literature (Zhu et al. 2012; 2008). Our model, however, contains side-way pairwise terms which form closed loops. These closed loops are fairly small so DP is still possible. In this paper, we combine the DP algorithm with state pruning for model inference, which has a bottom-up scoring step and a top-down backtracking step.

For the bottom-up scoring step, we compute the score of each vertex (to be specific, the score of the subgraph rooted at that vertex) in a bottom-up manner, only retraining the top-k scored candidate state configurations of each vertex for subsequent inference. The score of a subgraph $\mathcal{Q} \subseteq \mathcal{G}$ $(\mathcal{Q} = (\mathcal{V}_Q, \mathcal{E}_Q))$ is defined in Equ. (11), which is equivalent to Equ. (5) when $\mathcal{Q} = \mathcal{G}$.

$$F_Q(\mathbf{Y}_Q,\mathbf{Z}_Q|\tilde{\mathcal{S}},\mathcal{L},\mathcal{H}) = \sum_{p\in\mathcal{T}\cap\mathcal{V}_Q} f(y_p,z_p) + \sum_{c\in\mathcal{N}\cap\mathcal{V}_Q} f(y_c,z_c,\{(y_\mu,z_\mu):\mu\in Ch(c,z_c)\}) \qquad (11)$$

We set $k = 10$, making the inference procedure tractable with a moderate number of state configurations for each vertex. We show in a diagnostic experiment that this greedy pruning scarcely affects the quality of the result while reducing the inference time significantly.

After getting the score of the root vertex (i.e., whole human body), we backtrack the optimum state value from the retained top-k list for each vertex in a top-down manner. Concretely, for each part composition vertex c we select the best scored state configuration value of $(y_c, z_c, \{(y_\mu, z_\mu) : \mu \in Ch(c, z_c)\})$, and recursively infer the optimum state values of the selected child vertices given each $\mu \in Ch(c, z_c)$ as the root vertex of a subgraph. In the end, we can obtain the best parse tree from the pruned solution space of our AOG, and output corresponding state values $(\mathbf{Y},\mathbf{Z})$ to produce the final parsing result.

Experiments

Data & implementation details. We evaluate our algorithm on the Penn-Fudan benchmark (Wang et al. 2007), which consists of pedestrians in outdoor scenes with much pose variation. Because this dataset only provides testing data, following previous works (Bo and Fowlkes 2011; Rauschert and Collins 2012; Luo, Wang, and Tang 2013), we

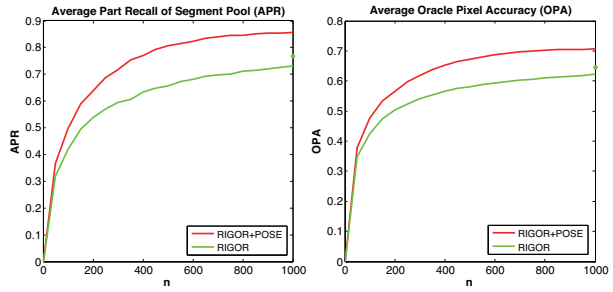


Figure 8: Comparison of our part segment proposal method (RIGOR+POSE) to the baseline (RIGOR). The green asterisks on the plots represent the APR/AOI of the RIGOR pool for the pool size $n = 2000$. (Best viewed in color)

train our parsing models using the HumanEva dataset (Sigal and Black 2006), which contains 937 images with pixel-level label maps for parts annotated by Bo and Fowlkes. The labels of the two datasets are consistent, which include 7 body parts { hair, face, upper-clothes, lower-clothes, arms (arm skin), legs (leg skin), and shoes }. For the pose model, we use the model provided by Chen and Yuille, trained on the Leeds Sports Pose Dataset (Johnson and Everingham 2010).

In part ranking & selection, we train linear SVR models for $P = 10$ part categories and select top $n_p = 10$ segments for each part category, as candidates of the final assembling stage. We treat left part and right part as two different part categories. For the segment feature used in the AOG (i.e. the unary term), we first normalize each kind of feature independently, then concatenate them together and normalize the whole feature. All the normalization is done with $L2$ norm. For simplicity, we only train one SVR model $g^p(s_i|\mathcal{L}, \mathcal{H})$ for each part category p so that $g^p_{z_p} = g^p, \forall z_p \neq 0$ in Equ. (6). However, due to the weight parameter $w^p_{z_p}$ is dependent on the part type z_p in training the AOG, the unary terms of different part subtypes are type-specific in the AOG model.

Effectiveness of pose for part proposal generation. We first investigate how the pose cues help the part proposal generation. Specifically, we compare our pose-guided segment proposal method with the baseline algorithm, i.e. the standard RIGOR algorithm (Humayun, Li, and Rehg 2014).

For evaluating the proposal algorithms, two standard criteria are used, i.e. average part recall (APR) and average part oracle IoU (AOI). The first measures how much portion of the ground-truth segments is covered (i.e., over 50% IoU) by the proposals, and the second measures the best IoU the proposal pool can achieve on average of all ground-truth segments. As shown in Fig. 8, our method significantly improves the quality of part segment proposals compared to the baseline by over 10% on average.

Effectiveness of features for part proposal selection. To investigate various features and their complementary properties for part proposal selection, we sequentially add them into our SVR model and test the performance/quality of the selected part segments.

In Tab. 2, we report the AOI scores for the top-1 ranked part segment and the top-10 ranked part segments respectively. Firstly, we can see the performance monotonically improves with more features used, which demonstrates the effectiveness of all features we proposed. By comparing (2) and (3), we can see a significant boost of the top-1 accuracy, indicating that the pose information becomes much more effective with the coded pose context feature. Finally, by adding the deep semantic feature in (4), the performance of selected part segment improves further. We set $n_p = 10$ because it strikes a good trade off between the quality and pool size of the selected part segments.

Methods	hair	face	u-cloth	l-cloth	arms	legs	shoes	mean
(1): *o2p + skin*	57.1	53.5	70.9	70.9	26.6	20.4	15.6	45.0
	68.8	66.9	80.0	81.4	54.6	55.3	45.3	64.6
(2): (1) + *pose*	61.7	58.6	73.2	72.7	29.9	23.4	17.5	48.1
	69.9	66.4	80.6	82.3	56.4	54.3	45.8	65.1
(3): (2) + *c-pose*	61.8	58.9	73.2	71.9	39.8	44.8	26.5	53.8
	69.9	66.4	80.5	82.4	55.8	59.1	47.4	65.9
(4): (3) + *fcn*	64.4	59.0	77.4	77.1	41.4	43.6	35.1	56.9
	70.7	66.6	82.2	83.4	55.9	59.3	48.8	66.7

Table 2: Comparison of four part models by AOI score (%) for top-1 ranked segment (top) and top-10 ranked segments (bottom). Models are numbered as (1) to (4), from top to bottom.

Effectiveness of the AOG. To show the effectiveness of our AOG design, we set up two experimental baselines for comparison: (1) Naive Assembling: considering only the unary terms and basic geometric constraints as defined in the paper (Bo and Fowlkes 2011), e.g. upper-clothes and lower-clothes must be adjacent. (2) Basic AOG: considering only the unary terms and the parent-child pairwise terms, without the side-way pairwise terms.

Tab. 3 shows that the basic AOG with parent-child spatial relations outperforms the naive assembling model, and by adding the pairwise side-way edges, the performance boosts further, which demonstrates the effectiveness of each component in our AOG model. For comparison, we also test the result of the AOG model without state pruning, which clearly justifies the use of state pruning in AOG inference. We can see that state pruning leads to neglectable decrease in accuracy while it reduces the inference runtime significantly, from 2 min. to 1 sec. per image.

Comparisons to the state of the art. We compare our approach with four state-of-the-art methods in literature, namely FCN (Wang et al. 2015), SBP (Bo and Fowlkes 2011), P&S (Rauschert and Collins 2012), and DDN (Luo, Wang, and Tang 2013). Specially, for FCN, we use the code provided by Wang et al. (2015) and re-train the networks with our training set.

Methods	hair	face	u-cloth	arms	l-cloth	legs	Avg
Naive Assembling	62.3	53.5	77.8	36.9	78.3	28.2	56.2
Basic AOG	63.1	52.9	77.1	38.0	78.1	35.9	57.5
Ours	**63.2**	**56.2**	**78.1**	**40.1**	**80.0**	45.5	**60.5**
Ours (w/o pruning)	63.2	56.2	78.1	40.1	80.0	**45.8**	60.5

Table 3: Per-pixel accuracy (%) of our AOG and two baselines.

Method	hair	face	u-cloth	arms	l-cloth	legs	shoes	Avg*
FCN	48.7	49.1	70.2	33.9	69.6	29.9	**36.1**	50.2
P&S	40.0	42.8	75.2	24.7	73.0	46.6	-	50.4
SBP	44.9	**60.8**	74.8	26.2	71.2	42.0	-	53.3
DDN	43.2	57.1	77.5	27.4	75.3	**52.3**	-	56.2
Ours	**63.2**	56.2	**78.1**	**40.1**	**80.0**	45.5	35.0	**60.5**

Table 4: Comparison of our approach with other state-of-the-art methods on the Penn-Fudan dataset in terms of per-pixel accuracy (%). The Avg* means the average without shoes class since it was not reported in other methods.

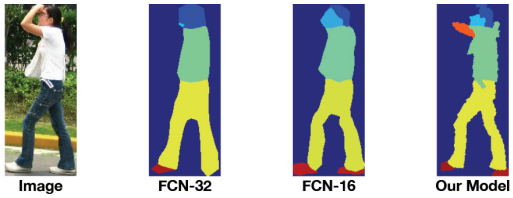
We show the quantitative comparison in Tab. 4. Our model outperforms the FCN by over 10% and the state-of-the-art DDN method by over 4%, from which we can see most improvement is from small parts such as hair and arms. It implies that by using the pose cues we can produce high-quality segment candidates that align to the boundaries for small parts. In addition, our AOG model together with the pose-context feature can leverage long-range spatial context information, making our model robust in shape variations and appearance ambiguities.

Fig. 9 (a) illustrates that our model is better at parsing small parts than FCN-32s and FCN-16s; Fig. 9 (b) gives typical parsing results of our method on Penn-Fudan; Fig. 9 (c) shows three failure examples due to color confusion with other objects, multiple instance occlusion, and large variation in lighting respectively, which generally fail most of current human parsing systems. For the first and the third failure cases, we got accurate pose estimation but failed to generate satisfactory segment proposals for lower-clothes, which suggests that we either adopt stronger shape cues in the segment proposal stage or seek richer context information (e.g. handbag in the first case). For the second case, we got a bad pose estimation due to occlusion and thus mixed two people's parts during assembling, which indicates the necessity of handling instance-level pose estimation or segmentation.

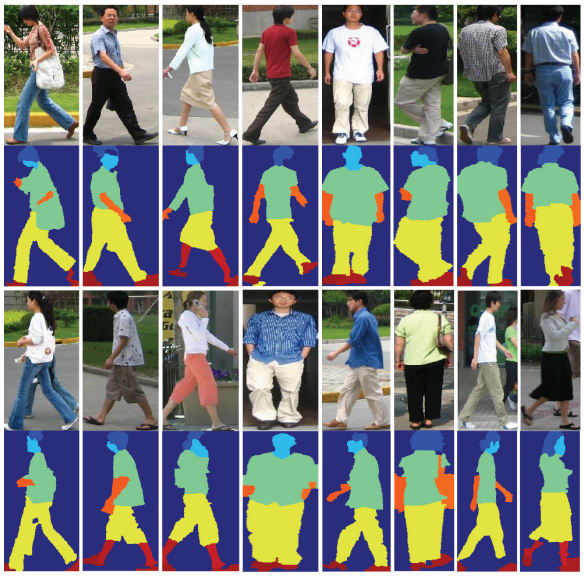
Conclusion

The contributions of this work are threefold: (1) We present an AOG-based human parsing pipeline which integrates top-down pose information into all three stages (i.e. pose-guided part proposals and pose-context features in part selection & assembling), obtaining state-of-the-art segmentation accuracy on a benchmark human parsing dataset, Penn-Fudan; (2) We propose semantically meaningful pose-context features that describe the geometric relationship between segment and pose joints; (3) We present a modular AOG with flexible composition structure. We show extensive experimental results that validate the effectiveness of each component of our pipeline.

In the future, we may adopt useful shape cues for the part proposal and selection stages, and combine CNN with graphical models in a more efficient way to better utilize their complementary role in the human parsing task.



(a) Comparision between our method and FCN.



(b) Additional parsing results of our method.

(c) Some failure cases of our method.

Figure 9: Qualitative results of our method on the Penn-Fudan dataset.

Acknowledgments

We would like to gratefully acknowledge support from NSF Expedion in Computer Science "Visual Cortex on Silicon" with award CCF-1317376, and ONR N00014-15-1-2356. We also thank NVIDIA for providing us with free GPUs that are used to train deep models.

References

Bo, Y., and Fowlkes, C. C. 2011. Shape-based pedestrian parsing. In *CVPR*.

Carreira, J. a.; Caseiro, R.; Batista, J.; and Sminchisescu, C. 2012. Semantic segmentation with second-order pooling. In *ECCV*.

Chen, X., and Yuille, A. 2014. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *NIPS*.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. 2015. Semantic image segmentation with deep convolutional nets and fully connected crfs. In *ICLR*.

Cheng, D. S.; Cristani, M.; Stoppa, M.; Bazzani, L.; and Murino, V. 2011. Custom pictorial structures for re-identification. In *BMVC*.

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

Hariharan, B.; Arbeláez, P.; Girshick, R.; and Malik, J. 2015. Hypercolumns for object segmentation and fine-grained localization. In *CVPR*.

Humayun, A.; Li, F.; and Rehg, J. M. 2014. RIGOR: Reusing Inference in Graph Cuts for generating Object Regions. In *Computer Vision and Pattern Recognition (CVPR), Proceedings of IEEE Conference on*. IEEE.

Joachims, T.; Finley, T.; and Yu, C.-N. J. 2009. Cutting-plane training of structural svms. *Machine Learning* 77(1):27–59.

Johnson, S., and Everingham, M. 2010. Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the British Machine Vision Conference*. doi:10.5244/C.24.12.

Khan, R.; Hanbury, A.; ; and Stottinger, J. 2010. Skin detection: A random forest approach. In *ICIP*.

Liu, S.; Liang, X.; Liu, L.; Shen, X.; Yang, J.; Xu, C.; Lin, L.; Cao, X.; and Yan, S. 2015. Matching-cnn meets knn: Quasi-parametric human parsing. In *CVPR*.

Liu, L.; Wang, L.; and Liu, X. 2011. In defense of soft-assignment coding. In *ICCV*.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. *CVPR*.

Luo, P.; Wang, X.; and Tang, X. 2013. Pedestrian parsing via deep decompositional network. In *ICCV*.

Ma, L.; Yang, X.; Xu, Y.; and Zhu, J. 2011. Human identification using body prior and generalized emd. In *ICIP*.

Rauschert, I., and Collins, R. T. 2012. A generative model for simultaneous estimation of human body shape and pixel-level segmentation. In *ECCV*.

Sigal, L., and Black, M. J. 2006. Synchronized video and motion capture dataset for evaluation of articulated human motion. *Technical Report CS-06-08, Brown University*.

Tsogkas, S.; Kokkinos, I.; Papandreou, G.; and Vedaldi, A. 2015. Semantic part segmentation with deep learning. *arXiv preprint arXiv:1505.02438*.

Wang, J., and Yuille, A. 2015. Semantic part segmentation using compositional model combining shape and appearance. In *CVPR*.

Wang, L.; Shi, J.; Song, G.; and Shen, I. F. 2007. Object detection combining recognition and segmentation. In *ACCV*.

Wang, J.; Yang, J.; Yu, K.; Lv, F.; Huang, T.; and Gong, Y. 2010. Locality-constrained linear coding for image classification. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 3360–3367. IEEE.

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

Wang, P.; Shen, X.; Lin, Z.; Cohen, S.; Price, B.; and Yuille, A. 2015. Joint object and part segmentation using deep learned potentials. *arXiv preprint arXiv:1505.00276*.

Wang, C.; Wang, Y.; and Yuille, A. L. 2013. An approach to pose-based action recognition. In *CVPR*.

Weber, M.; Bauml, M.; and Stiefelhagen, R. 2011. Part-based clothing segmentation for person retrieval. In *AVSS*.

Yadollahpour, P.; Batra, D.; and Shakhnarovich, G. 2013. Discriminative re-ranking of diverse segmentations. In *CVPR*.

Yamaguchi, K.; Luis, M. H. K.; Ortiz, E.; and Berg, T. L. 2012. Parsing clothing in fashion photographs. In *CVPR*.

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

Yang, M., and Yu, K. 2011. Real-time clothing recognition in surveillance videos. In *ICIP*.

Yang, J.; Yu, K.; Gong, Y.; and Huang, T. 2009. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 1794–1801. IEEE.

Yang, W.; Luo, P.; and Lin, L. 2014. Clothing co-parsing by joint image segmentation and labeling. In *CVPR*.

Zhu, S.-C., and Mumford, D. 2007. A stochastic grammar of images. *Foundations and Trends in Computer Graphics and Vision* 2(4):259–362.

Zhu, L.; Chen, Y.; Lu, Y.; Lin, C.; and Yuille, A. 2008. Max margin and/or graph learning for parsing the human body. In *CVPR*.

Zhu, J.; Wu, T.; Zhu, S.-C.; Yang, X.; and Zhang, W. 2012. Learning reconfigurable scene representation by tangram model. In *WACV*.

Zhu, J.; Wang, B.; Yang, X.; Zhang, W.; and Tu, Z. 2013. Action recognition with actons. In *ICCV*.