DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation

Leonid Pishchulin[1], Eldar Insafutdinov[1], Siyu Tang[1], Bjoern Andres[1],
Mykhaylo Andriluka[1,3], Peter Gehler[2], and Bernt Schiele[1]

[1]Max Planck Institute for Informatics, Germany
[2]Max Planck Institute for Intelligent Systems, Germany
[3]Stanford University, USA

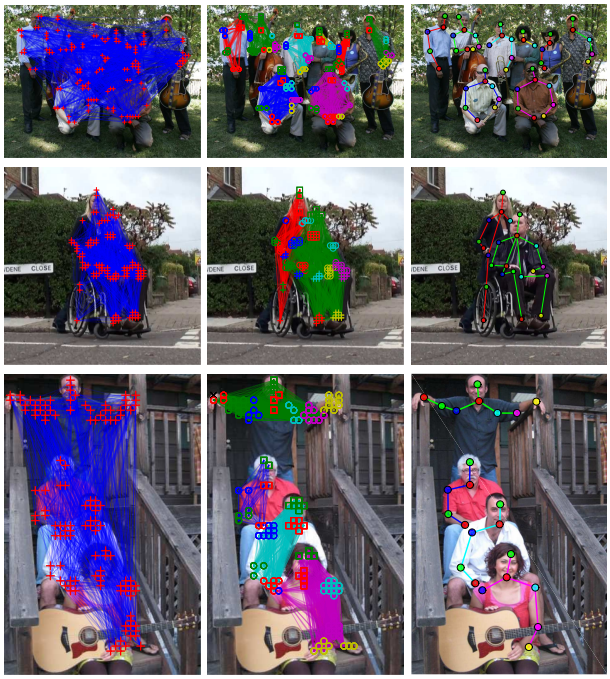
Abstract

This paper considers the task of articulated human pose estimation of multiple people in real world images. We propose an approach that jointly solves the tasks of detection and pose estimation: it infers the number of persons in a scene, identifies occluded body parts, and disambiguates body parts between people in close proximity of each other. This joint formulation is in contrast to previous strategies, that address the problem by first detecting people and subsequently estimating their body pose. We propose a partitioning and labeling formulation of a set of body-part hypotheses generated with CNN-based part detectors. Our formulation, an instance of an integer linear program, implicitly performs non-maximum suppression on the set of part candidates and groups them to form configurations of body parts respecting geometric and appearance constraints. Experiments on four different datasets demonstrate state-of-the-art results for both single person and multi person pose estimation[1].

1. Introduction

Human body pose estimation methods have become increasingly reliable. Powerful body part detectors [29] in combination with tree-structured body models [30, 7] show impressive results on diverse datasets [18, 3, 26]. These benchmarks promote pose estimation of single pre-localized persons but exclude scenes with multiple people. This problem definition has been a driver for progress, but also falls short on representing a realistic sample of real-world images. Many photographs contain multiple people of interest (see Fig 1) and it is unclear whether single pose approaches generalize directly. We argue that the multi person case deserves more attention since it is an important real-world task.

Key challenges inherent to multi person pose estimation are the partial visibility of some people, significant overlap of bounding box regions of people, and the a-priori unknown number of people in an image. The problem thus is to infer the number of persons, assign part detections to person instances while respecting geometric and appearance constraints. Most strategies use a two-stage inference process [23, 15, 28] to first detect and then independently estimate poses. This is unsuited for cases when people are in close



(a) (b) (c)

Figure 1. Method overview: (a) initial detections (= part candidates) and pairwise terms (graph) between all detections that (b) are jointly clustered belonging to one person (one colored subgraph = one person) and each part is labeled corresponding to its part class (different colors and symbols correspond to different body parts); (c) shows the predicted pose sticks.

[1]Models and code available at http://pose.mpi-inf.mpg.de

proximity since they permit simultaneous assignment of the same body-part candidates to multiple people hypotheses.

As a principled solution for multi person pose estimation a model is proposed that jointly estimates poses of all people present in an image by minimizing a joint objective. The formulation is based on partitioning and labeling an initial pool of body part candidates into subsets that correspond to sets of mutually consistent body-part candidates and abide to mutual consistency and exclusion constraints. The proposed method has a number of appealing properties. (1) The formulation is able to deal with an unknown number of people, and also infers this number by linking part hypotheses. (2) The formulation allows to either deactivate or merge part hypotheses in the initial set of part candidates hence effectively performing non-maximum suppression (NMS). In contrast to NMS performed on individual part candidates, the model incorporates evidence from all other parts making the process more reliable. (3) The problem is cast in the form of an Integer Linear Program (ILP). Although the problem is NP-hard, the ILP formulation facilitates the computation of bounds and feasible solutions with a certified optimality gap.

This paper makes the following contributions. The main contribution is the derivation of a joint detection and pose estimation formulation cast as an integer linear program. Further, two CNN variants are proposed to generate representative sets of body part candidates. These, combined with the model, obtain state-of-the-art results for both single-person and multi-person pose estimation on different datasets.

Related work. Most work on pose estimation targets the single person case. Methods progressed from simple part detectors and elaborate body models [25, 24, 16] to tree-structured pictorial structures (PS) models with strong part detectors [22, 34, 7, 26]. Impressive results are obtained predicting locations of parts with convolutional neural networks (CNN) [31, 29]. While body models are not a necessary component for effective part localization, constraints among parts allow to assemble independent detections into body configurations as demonstrated in [7] by combining CNN-based body part detectors with a body model [34].

A popular approach to multi-person pose estimation is to detect people first and then estimate body pose independently [28, 23, 34, 15]. [34] proposes a flexible mixture-of-parts model for detection and pose estimation. [34] obtains multiple pose hypotheses corresponding to different root part positions and then performing non-maximum suppression. [15] detects people using a flexible configuration of poselets and the body pose is predicted as a weighted average of activated poselets. [23] detects people and then predicts poses of each person using a PS model. [5] estimates poses of multiple people in 3D by constructing a shared space of 3D body part hypotheses, but uses 2D person detections to establish the number of people in the scene. These approaches are limited to cases with people sufficiently far from each other that do not have overlapping body parts.

Our work is closely related to [12, 21] who also propose a joint objective to estimate poses of multiple people. [12] proposes a multi-person PS model that explicitly models depth ordering and person-person occlusions. Our formulation is not limited by a number of occlusion states among people. [21] proposes a joint model for pose estimation and body segmentation coupling pose estimates of individuals by image segmentation. [12, 21] uses a person detector to generate initial hypotheses for the joint model. [21] resorts to a greedy approach of adding one person hypothesis at a time until the joint objective can be reduced, whereas our formulation can be solved with a certified optimality gap. In addition [21] relies on expensive labeling of body part segmentation, which the proposed approach does not require.

Similarly to [8] we aim to distinguish between visible and occluded body parts. [8] primarily focuse on the single-person case and handles multi-person scenes akin to [34]. We consider the more difficult problem of full-body pose estimation, whereas [12, 8] focus on upper-body poses and consider a simplified case of people seen from the front.

Our work is related to early work on pose estimation that also relies on integer linear programming to assemble candidate body part hypotheses into valid configurations [16]. Their single person method employs a tree graph augmented with weaker non-tree repulsive edges and expects the same number of parts. In contrast, our novel formulation relies on fully connected model to deal with unknown number of people per image and body parts per person.

The Minimum Cost Multicut Problem [9, 11], known in machine learning as correlation clustering [4], has been used in computer vision for image segmentation [1, 2, 19, 35] but has not been used before in the context of pose estimation. It is known to be NP-hard [10].

2. Problem Formulation

In this section, the problem of estimating articulated poses of an unknown number of people in an image is cast as an optimization problem. The goal of this formulation is to state three problems jointly: 1. The selection of a subset of body parts from a set D of *body part candidates*, estimated from an image as described in Section 4 and depicted as nodes of a graph in Fig. 1(a). 2. The *labeling* of each selected body part with one of C *body part classes*, e.g., "arm", "leg", "torso", as depicted in Fig. 1(c). 3. The *partitioning* of body parts that belong to the same person, as depicted in Fig. 1(b).

2.1. Feasible Solutions

We encode labelings of the three problems jointly through triples (x, y, z) of binary random variables with domains $x \in \{0,1\}^{D\times C}, y \in \{0,1\}^{\binom{D}{2}}$ and $z \in \{0,1\}^{\binom{D}{2}\times C^2}$. Here, $x_{dc} = 1$ indicates that body part candidate d is of

class c, $y_{dd'} = 1$ indicates that the body part candidates d and d' belong to the same person, and $z_{dd'cc'}$ are auxiliary variables to relate x and y through $z_{dd'cc'} = x_{dc}x_{d'c'}y_{dd'}$. Thus, $z_{dd'cc'} = 1$ indicates that body part candidate d is of class c ($x_{dc} = 1$), body part candidate d' is of class c' ($x_{d'c'} = 1$), and body part candidates d and d' belong to the same person ($y_{dd'} = 1$).

In order to constrain the 01-labelings (x, y, z) to well-defined articulated poses of one or more people, we impose the linear inequalities (1)–(3) stated below. Here, the inequalities (1) guarantee that every body part is labeled with at most one body part class. (If it is labeled with no body part class, it is suppressed). The inequalities (2) guarantee that distinct body parts d and d' belong to the same person only if neither d nor d' is suppressed. The inequalities (3) guarantee, for any three pairwise distinct body parts, d, d' and d'', if d and d' are the same person (as indicated by $y_{dd'} = 1$) and d' and d'' are the same person (as indicated by $y_{d'd''} = 1$), then also d and d'' are the same person ($y_{dd''} = 1$), that is, transitivity, cf. [9]. Finally, the inequalities (4) guarantee, for any $dd' \in \binom{D}{2}$ and any $cc' \in C^2$ that $z_{dd'cc'} = x_{dc}x_{d'c'}y_{dd'}$. These constraints allow us to write an objective function as a linear form in z that would otherwise be written as a cubic form in x and y. We denote by X_{DC} the set of all (x, y, z) that satisfy all inequalities, i.e., the set of feasible solutions.

$$\forall d \in D \forall cc' \in \tbinom{C}{2}: \quad x_{dc} + x_{dc'} \leq 1 \tag{1}$$

$$\begin{aligned} \forall dd' \in \tbinom{D}{2}: \quad & y_{dd'} \leq \sum_{c \in C} x_{dc} \\ & y_{dd'} \leq \sum_{c \in C} x_{d'c} \end{aligned} \tag{2}$$

$$\forall dd'd'' \in \tbinom{D}{3}: \quad y_{dd'} + y_{d'd''} - 1 \leq y_{dd''} \tag{3}$$

$$\begin{aligned} \forall dd' \in \tbinom{D}{2} \forall cc' \in C^2: \quad & x_{dc} + x_{d'c'} + y_{dd'} - 2 \leq z_{dd'cc'} \\ & z_{dd'cc'} \leq x_{dc} \\ & z_{dd'cc'} \leq x_{d'c'} \\ & z_{dd'cc'} \leq y_{dd'} \end{aligned} \tag{4}$$

When at most one person is in an image, we further constrain the feasible solutions to a well-defined pose of a single person. This is achieved by an additional class of inequalities which guarantee, for any two distinct body parts that are not suppressed, that they must be clustered together:

$$\forall dd' \in \tbinom{D}{2} \forall cc' \in C^2: \quad x_{dc} + x_{d'c'} - 1 \leq y_{dd'} \tag{5}$$

2.2. Objective Function

For every pair $(d, c) \in D \times C$, we will estimate a probability $p_{dc} \in [0, 1]$ of the body part d being of class c. In the context of CRFs, these probabilities are called *part unaries* and we will detail their estimation in Section 4.

For every $dd' \in \binom{D}{2}$ and every $cc' \in C^2$, we consider a probability $p_{dd'cc'} \in (0, 1)$ of the conditional probability of d and d' belonging to the same person, given that d and d' are body parts of classes c and c', respectively. For $c \neq c'$, these probabilities $p_{dd'cc'}$ are the *pairwise terms* in a graphical model of the human body. In contrast to the classic pictorial structures model, our model allows for a *fully connected graph* where each body part is connected to all other parts in the entire set D by a pairwise term. For $c = c'$, $p_{dd'cc'}$ is the probability of the part candidates d and d' representing the same part of the same person. This facilitates *clustering* of multiple part candidates of the same part of the same person and a *repulsive* property that prevents nearby part candidates of the same type to be associated to different people.

The optimization problem that we call the *subset partition and labeling problem* is the ILP that minimizes over the set of feasible solutions X_{DC}:

$$\min_{(x,y,z) \in X_{DC}} \langle \alpha, x \rangle + \langle \beta, z \rangle, \tag{6}$$

where we used the short-hand notation

$$\alpha_{dc} := \log \frac{1 - p_{dc}}{p_{dc}} \tag{7}$$

$$\beta_{dd'cc'} := \log \frac{1 - p_{dd'cc'}}{p_{dd'cc'}} \tag{8}$$

$$\langle \alpha, x \rangle := \sum_{d \in D} \sum_{c \in C} \alpha_{dc}\, x_{dc} \tag{9}$$

$$\langle \beta, z \rangle := \sum_{dd' \in \binom{D}{2}} \sum_{c,c' \in C} \beta_{dd'cc'}\, z_{dd'cc'} \quad . \tag{10}$$

The objective (6)–(10) is the MAP estimate of a probability measure of joint detections x and clusterings y, z of body parts, where prior probabilities p_{dc} and $p_{dd'cc'}$ are estimated *independently* from data, and the likelihood is a positive constant if (x, y, z) satisfies (1)–(4), and is 0, otherwise. The exact form (6)–(10) is obtained when minimizing the negative logarithm of this probability measure.

2.3. Optimization

In order to obtain feasible solutions of the ILP (6) with guaranteed bounds, we separate the inequalities (1)–(5) in the branch-and-cut loop of the state-of-the-art ILP solver Gurobi. More precisely, we solve a sequence of relaxations of the problem (6), starting with the (trivial) unconstrained problem. Each problem is solved using the cuts proposed by Gurobi. Once an integer feasible solution is found, we identify violated inequalities (1)–(5), if any, by breadth-first-search, add these to the constraint pool and re-solve the tightened relaxation. Once an integer solution satisfying all inequalities is found, together with a lower bound that certifies an optimality gap below 1%, we terminate.

3. Pairwise Probabilities

Here we describe the estimation of the pairwise terms. We define pairwise features $f_{dd'}$ for the variable $z_{dd'cc'}$

(Sec. 2). Each part detection d includes the probabilities $f_{p_{dc}}$ (Sec. 4.4), its location (x_d, y_d), scale h_d and bounding box B_d coordinates. Given two detections d and d', and the corresponding features $(f_{p_{dc}}, x_d, y_d, h_d, B_d)$ and $(f_{p_{d'c}}, x_{d'}, y_{d'}, h_{d'}, B_{d'})$, we define two sets of auxiliary variables for $z_{dd'cc'}$, one set for $c = c'$ (same body part class clustering) and one for $c \neq c'$ (across two body part classes labeling). These features capture the proximity, kinematic relation and appearance similarity between body parts.

The same body part class ($c = c'$). Two detections denoting the same body part of the same person should be in close proximity to each other. We introduce the following auxiliary variables that capture the spatial relations: $\Delta x = |x_d - x_{d'}|/\bar{h}$, $\Delta y = |y_d - y_{d'}|/\bar{h}$, $\Delta h = |h_d - h_{d'}|/\bar{h}$, $IOUnion$, $IOMin$, $IOMax$. The latter three are intersections over union/minimum/maximum of the two detection boxes, respectively, and $\bar{h} = (h_d + h_{d'})/2$.

Non-linear Mapping. We augment the feature representation by appending quadratic and exponential terms. The final pairwise feature $f_{dd'}$ for the variable $z_{dd'cc}$ is $(\Delta x, \Delta y, \Delta h, IOUnion, IOMin, IOMax, (\Delta x)^2, \ldots, (IOMax)^2, \exp(-\Delta x), \ldots, \exp(-IOMax))$.

Two different body part classes ($c \neq c'$). We encode the kinematic body constraints into the pairwise feature by introducing auxiliary variables $S_{dd'}$ and $R_{dd'}$, where $S_{dd'}$ and $R_{dd'}$ are the Euclidean distance and the angle between two detections, respectively. To capture the joint distribution of $S_{dd'}$ and $R_{dd'}$, instead of using $S_{dd'}$ and $R_{dd'}$ directly, we employ the posterior probability $p(z_{dd'cc'} = 1|S_{dd'}, R_{dd'})$ as pairwise feature for $z_{dd'cc'}$ to encode the geometric relations between the body part class c and c'. More specifically, assuming the prior probability $p(z_{dd'cc'} = 1) = p(z_{dd'cc'} = 0) = 0.5$, the posterior probability of detection d and d' have the body part label c and c', namely $z_{dd'cc'} = 1$, is

$$p(z_{dd'cc'} = 1|S_{dd'}, R_{dd'}) = \frac{p(S_{dd'}, R_{dd'}|z_{dd'cc'} = 1)}{p(S_{dd'}, R_{dd'}|z_{dd'cc'} = 1) + p(S_{dd'}, R_{dd'}|z_{dd'cc'} = 0)},$$

where $p(S_{dd'}, R_{dd'}|z_{dd'cc'} = 1)$ is obtained by conducting a normalized 2D histogram of $S_{dd'}$ and $R_{dd'}$ from positive training examples, analogous to the negative likelihood $p(S_{dd'}, R_{dd'}|z_{dd'cc'} = 0)$. In Sec. 5.1 we also experiment with encoding the appearance into the pairwise feature by concatenating the feature $f_{p_{dc}}$ from d and $f_{p_{d'c}}$ from d', as $f_{p_{dc}}$ is the output of the CNN-based part detectors. The final pairwise feature is $(p(z_{dd'cc'} = 1|S_{dd'}, R_{dd'}), f_{p_{dc}}, f_{p_{d'c}})$.

3.1. Probability Estimation

The coefficients α and β of the objective function (Eq. 6) are defined by the probability ratio in the log space (Eq. 7 and Eq. 8). Here we describe the estimation of the corresponding probability density: *(1)* For every pair of detection and part classes, namely for any $(d, c) \in D \times C$, we estimate a probability $p_{dc} \in (0, 1)$ of the detection d being a body part of class c. *(2)* For every combination of two distinct detections and two body part classes, namely for any $dd' \in \binom{D}{2}$ and any $cc' \in C^2$, we estimate a probability $p_{dd'cc'} \in (0, 1)$ of d and d' belonging to the same person, meanwhile d and d' are body parts of classes c and c', respectively.

Learning. Given the features $f_{dd'}$ and a Gaussian prior $p(\theta_{cc'}) = \mathcal{N}(0, \sigma^2)$ on the parameters, logistic model is

$$p(z_{dd'cc'} = 1|f_{dd'}, \theta_{cc'}) = \frac{1}{1 + \exp(-\langle \theta_{cc'}, f_{dd'} \rangle)}. \quad (11)$$

$(|C| \times (|C| + 1))/2$ parameters are estimated using ML.

Inference Given two detections d and d', the coefficients α_{dc} for x_{dc} and $\alpha_{d'c}$ for $x_{d'c}$ are obtained by Eq. 7, the coefficient $\beta_{dd'cc'}$ for $z_{dd'cc'}$ has the form

$$\beta_{dd'cc'} = \log \frac{1 - p_{dd'cc'}}{p_{dd'cc'}} = -\langle f_{dd'}, \theta_{cc'} \rangle. \quad (12)$$

Model parameters $\theta_{cc'}$ are learned using logistic regression.

4. Body Part Detectors

We first introduce our deep learning-based part detection models and then evaluate them on two prominent benchmarks thereby significantly outperforming state of the art.

4.1. Adapted Fast R-CNN (*AFR-CNN*)

To obtain strong part detectors we adapt Fast R-CNN [14]. FR-CNN takes as input an image and set of class-independent region proposals [32] and outputs the softmax probabilities over all classes and refined bounding boxes. To adapt FR-CNN for part detection we alter it in two ways: *1)* proposal generation and *2)* detection region size. The adapted version is called *AFR-CNN* throughout the paper.

Detection proposals. Generating object proposals is essential for FR-CNN, meanwhile detecting body parts is challenging due to their small size and high intra-class variability. We use DPM-based part detectors [22] for proposal generation. We collect K top-scoring detections by each part detector in a common pool of N part-independent proposals and use these proposals as input to *AFR-CNN*. N is $2,000$ in case of single and $20,000$ in case of multiple people.

Larger context. Increasing the size of DPM detections by upscaling every bounding box by a fixed factor allows to capture more context around each part. In Sec. 4.3 we evaluate the influence of upscaling and show that using larger context around parts is crucial for best performance.

Details. Following standard FR-CNN training procedure ImageNet models are finetuned on pose estimation task. Center of a predicted bounding box is used for body part location prediction. See supplemental for detailed parameter analysis.

Setting	Head	Sho	Elb	Wri	Hip	Knee	Ank	PCK	AUC
oracle 2,000	98.8	98.8	97.4	96.4	97.4	98.3	97.7	97.8	84.0
DPM scale 1	48.8	25.1	14.4	10.2	13.6	21.8	27.1	23.0	13.6
AlexNet scale 1	82.2	67.0	49.6	45.4	53.1	52.9	48.2	56.9	35.9
AlexNet scale 4	85.7	74.4	61.3	53.2	64.1	63.1	53.8	65.1	39.0
+ optimal params	88.1	79.3	68.9	62.6	73.5	69.3	64.7	72.4	44.6
VGG scale 4 optimal params	91.0	84.2	74.6	67.7	77.4	77.3	72.8	77.9	50.0
+ finetune LSP	**95.4**	**86.5**	**77.8**	**74.0**	**84.5**	**78.8**	**82.6**	**82.8**	**57.0**

Table 1. Unary only performance (PCK) of *AFR-CNN* on the LSP (Person-Centric) dataset. *AFR-CNN* is finetuned from ImageNet to MPII (lines 3-6), and then finetuned to LSP (line 7).

4.2. Dense Architecture (*Dense-CNN*)

Using proposals for body part detection may be suboptimal. We thus develop a fully convolutional architecture for computing part probability scoremaps.

Stride. We build on VGG [27]. Fully convolutional VGG has stride of 32 px – too coarse for precise part localization. We thus use hole algorithm [6] to reduce the stride to 8 px.

Scale. Selecting image scale is crucial. We found that scaling to a standing height of 340 px performs best: VGG receptive field sees entire body to disambiguate body parts.

Loss function. We start with a softmax loss that outputs probabilities for each body part and background. The downside is inability to assign probabilities above 0.5 to several close-by body parts. We thus re-formulate the detection as multi-label classification, where at each location a separate set of probability distributions is estimated for each part. We use sigmoid activation function on the output neurons and cross entropy loss. We found this loss to perform better than softmax and converge much faster compared to MSE [30]. Target training scoremap for each joint is constructed by assigning a positive label 1 at each location within 15 px to the ground truth, and negative label 0 otherwise.

Location refinement. In order to improve location precision we follow [14]: we add a location refinement FC layer after the FC7 and use the relative offsets $(\Delta x, \Delta y)$ from a scoremap location to the ground truth as targets.

Regression to other parts. Similar to location refinement we add an extra term to the objective function where for each part we regress onto all other part locations. We found this auxiliary task to improve the performance (c.f. Sec. 4.3).

Training. We follow best practices and use SGD for CNN training. In each iteration we forward-pass a single image. After FC6 we select all positive and random negative samples to keep the pos/neg ratio as 25%/75%. We finetune VGG from Imagenet model to pose estimation task and use training data augmentation. We train for 430k iterations with the following learning rates (lr): 10k at lr=0.001, 180k at lr=0.002, 120k at lr=0.0002 and 120k at lr=0.0001. Pretraining at smaller lr prevents the gradients from diverging.

Setting	Head	Sho	Elb	Wri	Hip	Knee	Ank	PCK	AUC
MPII softmax	91.5	85.3	78.0	72.4	81.7	80.7	75.7	80.8	51.9
+ LSPET	94.6	86.8	79.9	75.4	83.5	82.8	77.9	83.0	54.7
+ sigmoid	93.5	87.2	81.0	77.0	85.5	83.3	79.3	83.8	55.6
+ location refinement	95.0	88.4	81.5	76.4	88.0	83.3	80.8	84.8	61.5
+ auxiliary task	95.1	89.6	82.8	78.9	89.0	85.9	81.2	86.1	61.6
+ finetune LSP	**97.2**	**90.8**	**83.0**	**79.3**	**90.6**	**85.6**	**83.1**	**87.1**	**63.6**

Table 2. Unary only performance (PCK) of *Dense-CNN* VGG on LSP (PC) dataset. *Dense-CNN* is finetuned from ImageNet to MPII (line 1), to MPII+LSPET (lines 2-5), and finally to LSP (line 6).

4.3. Evaluation of Part Detectors

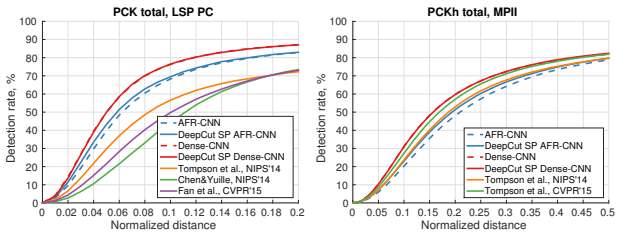
Datasets. We train and evaluate on three public benchmarks: "Leeds Sports Poses" (LSP) [17] (person-centric (PC)), "LSP Extended" (LSPET) [18][2], and "MPII Human Pose" ("Single Person") [3]. The MPII training set (19185 people) is used as default. In some cases LSP training *and* LSPET are added to MPII (marked as MPII+LSPET in the experiments).

Evaluation measures. We use the standard "PCK" metric [26, 31, 30] and evaluation scripts available on the web page of [3]. In addition, we report "Area under Curve" (AUC) computed for the entire range of PCK thresholds.

AFR-CNN. Evaluation of *AFR-CNN* on LSP is shown in Tab. 1. Oracle selecting per part the closest from $2,000$ proposals achieves 97.8% PCK, as proposals cover majority of the ground truth locations. Choosing a single proposal per part using DPM score achieves 23.0% PCK – not surprising given the difficulty of the body part detection problem. Rescoring the proposals using *AFR-CNN* with AlexNet [20] dramatically improves the performance to 56.9% PCK, as CNN learns richer image representations. Extending the regions by 4x (1x $\approx$ head size) achieves 65.1% PCK, as it incorporates more context including the information about symmetric parts and allows to implicitly encode higher-order part relations. Using data augmentation and slightly tuning training parameters improves the performance to 72.4% PCK. We refer to the supplementary material for detailed analysis. Deeper VGG architecture improves over smaller AlexNet reaching 77.9% PCK. All results so far are achieved by finetuning the ImageNet models on MPII. Further finetuning to LSP leads to remarkable 82.8% PCK: CNN learns LSP-specific image representations. Strong increase in AUC (57.0 vs. 50%) is due to improvements for smaller PCK thresholds. Using no bounding box regression leads to performance drop (81.3% PCK, 53.2% AUC): location refinement is crucial for better localization. Overall *AFR-CNN* obtains very good results on LSP by far outperforming the state of the art (c.f. Tab. 3, rows $7-9$). Evaluation on MPII shows competitive performance (Tab. 4, row 1).

Dense-CNN. The results are in Tab. 2. Training with VGG on MPII with softmax loss achieves 80.8% PCK thereby

[2]To reduce labeling noise we re-annotated original high-resolution images and make the data available at http://datasets.d2.mpi-inf.mpg.de/hr-lspet/hr-lspet.zip



(a) LSP (PC) (b) MPII Single Person

Figure 2. Pose estimation results over all PCK thresholds.

outperforming *AFR-CNN* (c.f. Tab. 1, row 6). This shows the advantages of fully convolutional training and evaluation. Expectedly, training on larger MPII+LSPET dataset improves the results (83.0 vs. 80.8% PCK). Using cross-entropy loss with sigmoid activations improves the results to 83.8% PCK, as it better models the appearance of close-by parts. Location refinement improves localization accuracy (84.8% PCK), which becomes more clear when analyzing AUC (61.5 vs. 55.6%). Interestingly, regressing to other parts further improves PCK to 86.1% showing a value of training with the auxiliary task. Finally, finetuning to LSP achieves the best result of 87.1% PCK, which is significantly higher than the best published results (c.f. Tab. 3, rows $7-9$). Unary-only evaluation on MPII reveals slightly higher AUC results compared to the state of the art (Tab. 4, row $3-4$).

4.4. Using Detections in DeepCut Models

The SPLP problem is NP-hard, to solve instances of it efficiently we select a subset of representative detections from the entire set produced by a model. In our experiments we use $|D| = 100$ as default detection set size. In case of the *AFR-CNN* we directly use the softmax output as unary probabilities: $f_{p_{dc}} = (p_{d1}, \ldots, p_{dc})$, where p_{dc} is the probability of the detection d being the part class c. For *Dense-CNN* detection model we use the sigmoid detection unary scores.

5. DeepCut Results

The aim of this paper is to tackle the multi person case. To that end, we evaluate the proposed *DeepCut* models on four diverse benchmarks. We confirm that both single person (*SP*) and multi person (*MP*) variants (Sec. 2) are effective on standard *SP* pose estimation datasets [17, 3]. Then, we demonstrate superior performance of *DeepCut MP* on the multi person pose estimation task.

5.1. Single Person Pose Estimation

We now evaluate single person (*SP*) and more general multi person (*MP*) *DeepCut* models on LSP and MPII *SP* benchmarks described in Sec. 4. Since this evaluation setting implicitly relies on the knowledge that all parts are present in the image we always output the full number of parts.

Results on LSP. We report per-part PCK results (Tab. 3) and results for a variable distance threshold (Fig. 2 (a)).

Setting	Head	Sho	Elb	Wri	Hip	Knee	Ank	PCK	AUC
AFR-CNN (unary)	95.4	86.5	77.8	74.0	84.5	82.6	78.8	82.8	57.0
+ *DeepCut SP*	95.4	86.7	78.3	74.0	84.3	82.9	79.2	83.0	58.4
+ appearance pairwise	95.4	87.2	78.6	73.7	84.7	82.8	78.8	83.0	58.5
+ *DeepCut MP*	95.2	86.7	78.2	73.5	84.6	82.8	79.0	82.9	58.0
Dense-CNN (unary)	**97.2**	90.8	83.0	**79.3**	90.6	85.6	**83.1**	**87.1**	**63.6**
+ *DeepCut SP*	97.0	91.0	**83.8**	78.1	91.0	86.7	82.0	**87.1**	63.5
+ *DeepCut MP*	96.2	**91.2**	83.3	77.6	**91.3**	**87.0**	80.4	86.7	62.6
Tompson et al. [30]	90.6	79.2	67.9	63.4	69.5	71.0	64.2	72.3	47.3
Chen&Yuille [7]	91.8	78.2	71.8	65.5	73.3	70.2	63.4	73.4	40.1
Fan et al. [33]*	92.4	75.2	65.3	64.0	75.7	68.3	70.4	73.0	43.2

* re-evaluated using the standard protocol, for details see project page of [33]

Table 3. Pose estimation results (PCK) on LSP (PC) dataset.

DeepCut SP AFR-CNN model using 100 detections improves over unary only (83.0 vs. 82.8% PCK, 58.4 vs. 57% AUC), as pairwise connections filter out some of the high-scoring detections on the background. The improvement is clear in Fig. 2 (a) for smaller thresholds. Using part appearance scores in addition to geometrical features in $c \neq c'$ pairwise terms only slightly improves AUC, as the appearance of neighboring parts is mostly captured by a relatively large region centered at each part. The performance of *DeepCut MP AFR-CNN* matches the *SP* and improves over *AFR-CNN* alone: *DeepCut MP* correctly handles the *SP* case. Performance of *DeepCut SP Dense-CNN* is almost identical to unary only, unlike the results for *AFR-CNN*. *Dense-CNN* performance is noticeably higher compared to *AFR-CNN*, and "easy" cases that could have been corrected by a spatial model are resolved by stronger part detectors alone.

Comparison to the state of the art (LSP). Tab. 3 compares results of *DeepCut* models to other deep learning methods specifically designed for single person pose estimation. All *DeepCuts* significantly outperform the state of the art, with *DeepCut SP Dense-CNN* model improving by 13.7% PCK over the best known result [7]. The improvement is even more dramatic for lower thresholds (Fig. 2 (a)): for PCK @ 0.1 the best model improves by 19.9% over Tompson et al. [30], by 26.7% over Fan et al. [33], and by 32.4% PCK over Chen&Yuille [7]. The latter is interesting, as [7] use a stronger spatial model that predicts the pairwise conditioned on the CNN features, whereas *DeepCuts* use geometric-only pairwise connectivity. Including body part orientation information into *DeepCuts* should further improve the results.

Results on MPII Single Person. Results are shown in Tab. 4 and Fig. 2 (b). *DeepCut SP AFR-CNN* noticeably improves over *AFR-CNN* alone (79.8 vs. 78.8% PCK, 51.1 vs. 49.0% AUC). The improvement is stronger for smaller thresholds (c.f. Fig. 2), as spatial model improves part localization. *Dense-CNN* alone trained on MPII outperforms *AFR-CNN* (81.6 vs. 78.8% PCK), which shows the advantages of dense training and evaluation. As expected, *Dense-CNN* performs slightly better when trained on the larger MPII+LSPET. Finally, *DeepCut Dense-CNN SP* is slightly better than *Dense-CNN* alone leading to the best

Setting	Head	Sho	Elb	Wri	Hip	Knee	Ank	PCK$_h$	AUC
AFR-CNN (unary)	91.5	89.7	80.5	74.4	76.9	69.6	63.1	78.8	49.0
+ *DeepCut SP*	92.3	90.6	81.7	74.9	79.2	70.4	63.0	79.8	51.1
Dense-CNN (unary)	93.5	88.6	82.2	77.1	81.7	74.4	**68.9**	81.6	56.0
+LSPET	94.0	89.4	82.3	77.5	82.0	74.4	68.7	81.9	**56.5**
+*DeepCut SP*	94.1	90.2	83.4	77.3	**82.6**	**75.7**	68.6	**82.4**	**56.5**
Tompson et al. [30]	95.8	90.3	80.5	74.3	77.6	69.7	62.8	79.6	51.8
Tompson et al. [29]	**96.1**	**91.9**	**83.9**	**77.8**	80.9	72.3	64.8	82.0	54.9

Table 4. Pose estimation results (PCK$_h$) on MPII Single Person.

Setting	Head	U Arms	L Arms	Torso	mPCP	AOP
AFR-CNN det ROI	69.8	46.0	36.7	83.7	53.1	73.9
DeepCut MP AFR-CNN	99.0	79.5	74.3	87.1	82.2	85.6
Dense-CNN det ROI	76.0	46.0	40.2	83.7	55.3	73.8
DeepCut MP Dense-CNN	**99.3**	**81.5**	**79.5**	87.1	**84.7**	**86.5**
Ghiasi et. al. [13]	-	-	-	-	63.6	74.0
Eichner&Ferrari [12]	97.6	68.2	48.1	86.1	69.4	80.0
Chen&Yuille [8]	98.5	77.2	71.3	**88.5**	80.7	84.9

Table 5. Pose estimation results (mPCP) on WAF dataset.

result on MPII dataset (82.4% PCK).

Comparison to the state of the art (MPII). We compare the performance of *DeepCut* models to the best deep learning approaches from the literature [30, 29][3]. *DeepCut SP Dense-CNN* outperforms both [30, 29] (82.4 vs 79.6 and 82.0% PCK, respectively). Similar to them *DeepCuts* rely on dense training and evaluation of part detectors, but unlike them use single size receptive field and do not include multi-resolution context information. Also, appearance and spatial components of *DeepCuts* are trained piece-wise, unlike [30]. We observe that performance differences are higher for smaller thresholds (c.f. Fig. 2 (b)). This is remarkable, as a much simpler strategy for location refinement is used compared to [29]. Using multi-resolution filters and joint training should improve the performance.

5.2. Multi Person Pose Estimation

We now evaluate *DeepCut MP* models on the challenging task of *MP* pose estimation with an unknown number of people per image and visible body parts per person.

Datasets. For evaluation we use two public *MP* benchmarks: "We Are Family" (WAF) [12] with 350 training and 175 testing group shots of people; "MPII Human Pose" ("Multi-Person") [3] consisting of 3844 training and 1758 testing groups of multiple interacting individuals in highly articulated poses with variable number of parts. On MPII, we use a subset of 288 testing images for evaluation. We first pre-finetune both *AFR-CNN* and *Dense-CNN* from ImageNet to MPII and MPII+LSPET, respectively, and further finetune each model to WAF and MPII Multi-Person. For WAF, we re-train the spatial model on WAF training set.

WAF evaluation measure. Approaches are evaluated using the official toolkit [12], thus results are directly comparable to prior work. The toolkit implements occlusion-aware "Percentage of Correct Parts (mPCP)" metric. In addition, we report "Accuracy of Occlusion Prediction (AOP)" [8].

MPII Multi-Person evaluation measure. PCK metric is suitable for *SP* pose estimation with known number of parts and does not penalize for false positives that are not a part of the ground truth. Thus, for *MP* pose estimation we use "Mean Average Precision (mAP)" measure, similar to [28, 34]. In contrast to [28, 34] evaluating the detection of *any* part instance in the image disrespecting inconsistent pose predictions, we evaluate consistent part configurations. First, multiple body pose predictions are generated and then assigned to the ground truth (GT) based on the highest PCK$_h$ [3]. Only single pose can be assigned to GT. Unassigned predictions are counted as false positives. Finally, AP for each body part is computed and mAP is reported.

Baselines. To assess the performance of *AFR-CNN* and *Dense-CNN* we follow a traditional route from the literature based on two stage approach: first a set of regions of interest (*ROI*) is generated and then the *SP* pose estimation is performed in the *ROIs*. This corresponds to unary only performance. *ROI* are either based on a ground truth (*GT ROI*) or on the people detector output (*det ROI*).

Results on WAF. Results are shown in Tab. 5. *det ROI* is obtained by extending provided upper body detection boxes. *AFR-CNN det ROI* achieves 57.6% mPCP and 73.9% AOP. *DeepCut MP AFR-CNN* significantly improves over *AFR-CNN det ROI* achieving 82.2% mPCP. This improvement is stronger compared to LSP and MPII due to several reasons. First, mPCP requires consistent prediction of body sticks as opposite to body joints, and including spatial model enforces consistency. Second, mPCP metric is occlusion-aware. *DeepCuts* can deactivate detections for the occluded parts thus effectively reasoning about occlusion. This is supported by strong increase in AOP (85.6 vs. 73.9%). Results by *DeepCut MP Dense-CNN* follow the same tendency achieving the best performance of 84.7% mPCP and 86.5% AOP. Both increase in mPCP and AOP show the advantages of *DeepCuts* over traditional *det ROI* approaches.

Tab. 5 shows that *DeepCuts* outperform all prior methods. Deep learning method [8] is outperformed both for mPCP (84.7 vs. 80.7%) and AOP (86.5 vs. 84.9%) measures. This is remarkable, as *DeepCuts* reason about part interactions across several people, whereas [8] primarily focuses on the single-person case and handles multi-person scenes akin to [34]. In contrast to [8], *DeepCuts* are not limited by the number of possible occlusion patterns and cover person-person occlusions and other types as truncation and occlusion by objects in one formulation. *DeepCuts* significantly outperform [12] while being more general: unlike [12] *DeepCuts* do not require person detector and not limited by a number of occlusion states among people.

Qualitative comparison to [8] is provided in Fig. 3.

[3] [30] was re-trained and evaluated on MPII dataset by the authors.

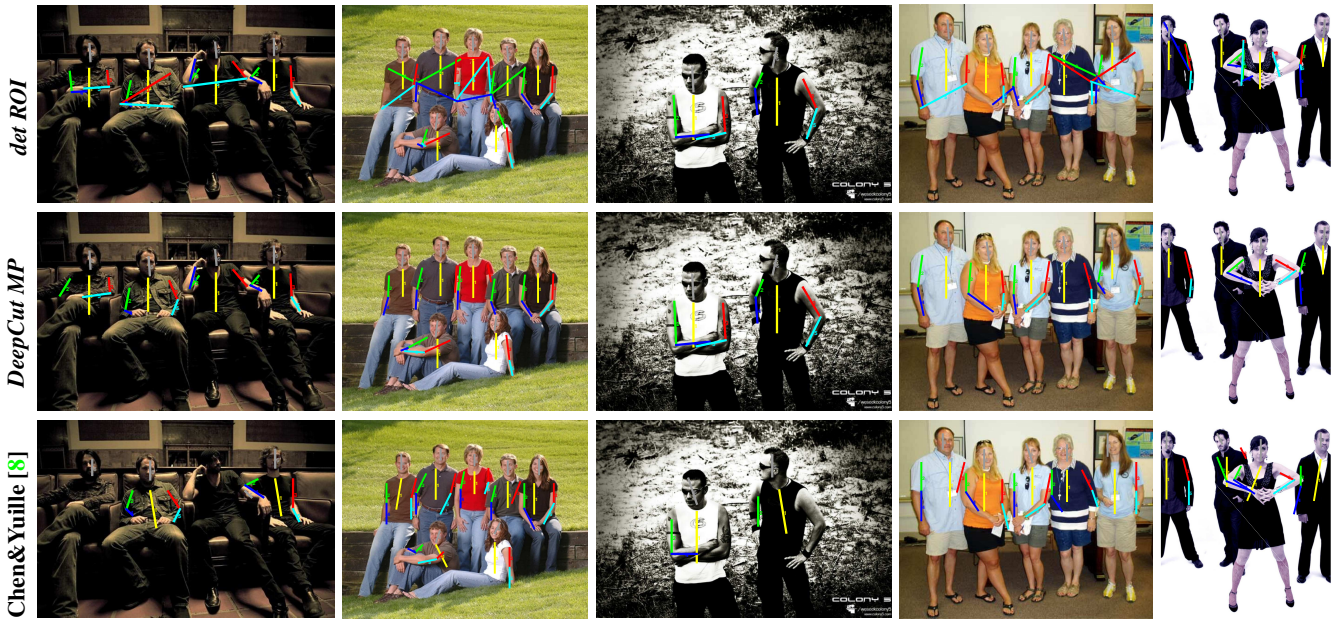


Figure 3. Qualitative comparison of our joint formulation *DeepCut MP Dense-CNN* (middle) to the traditional two-stage approach *Dense-CNN det ROI* (top) and the approach of Chen&Yuille [8] (bottom) on WAF dataset. In contrast to *det ROI*, *DeepCut MP* is able to disambiguate multiple and potentially overlapping persons and correctly assemble independent detections into plausible body part configurations. In contrast to [8], *DeepCut MP* can better predict occlusions (image 2 person $1-4$ from the left, top row; image 4 person 1, 4; image 5, person 2) and better cope with strong articulations and foreshortenings (image 1, person 1, 3; image 2 person 1 bottom row; image 3, person 1-2). See supplementary material for more examples.

Results on MPII Multi-Person. Obtaining a strong detector of highly articulated people having strong occlusions and truncations is difficult. We employ a neck detector as a person detector as it turned out to be the most reliable part. Full body bounding box is created around a neck detection and used as *det ROI*. *GT ROI*s were provided by the authors [3]. As the *MP* approach [8] is not public, we compare to *SP* state-of-the-art method [7] applied to *GT ROI* image crops.

Results are shown in Tab. 6. *DeepCut MP AFR-CNN* improves over *AFR-CNN det ROI* by 4.3% achieving 51.4% AP. The largest differences are observed for the ankle, knee, elbow and wrist, as those parts benefit more from the connections to other parts. *DeepCut MP UB AFR-CNN* using upper body parts only slightly improves over the full body model when compared on common parts (60.5 vs 58.2% AP). Similar tendencies are observed for *Dense-CNN*s, though improvements of *MP UB* over *MP* are more significant.

All *DeepCuts* outperform *Chen&Yuille SP GT ROI*, partially due to stronger part detectors compared to [7] (c.f. Tab. 3). Another reason is that *Chen&Yuille SP GT ROI* does not model body part occlusion and truncation always predicting the full set of parts, which is penalized by the AP measure. In contrast, our formulation allows to deactivate the part hypothesis in the initial set of part candidates thus effectively performing non-maximum suppression. In *DeepCuts* part hypotheses are suppressed based on the evidence from all other body parts making this process more reliable.

Setting	Head	Sho	Elb	Wri	Hip	Knee	Ank	UBody	FBody
AFR-CNN det ROI	71.1	65.8	49.8	34.0	47.7	36.6	20.6	55.2	47.1
AFR-CNN MP	71.8	67.8	54.9	38.1	52.0	41.2	30.4	58.2	51.4
AFR-CNN MP UB	75.2	71.0	56.4	39.6	-	-	-	60.5	-
Dense-CNN det ROI	77.2	71.8	55.9	42.1	53.8	39.9	27.4	61.8	53.2
Dense-CNN MP	73.4	71.8	57.9	39.9	**56.7**	**44.0**	**32.0**	60.7	**54.1**
Dense-CNN MP UB	**81.5**	**77.3**	**65.8**	**50.0**	-	-	-	**68.7**	-
AFR-CNN GT ROI	73.2	66.5	54.6	42.3	50.1	44.3	37.8	59.1	53.1
Dense-CNN GT ROI	78.1	74.1	62.2	52.0	56.9	48.7	46.1	66.6	60.2
Chen&Yuille SP GT ROI	65.0	34.2	22.0	15.7	19.2	15.8	14.2	34.2	27.1

Table 6. Pose estimation results (AP) on MPII Multi-Person.

6. Conclusion

Articulated pose estimation of multiple people in uncontrolled real world images is challenging but of real world interest. In this work, we proposed a new formulation as a joint subset partitioning and labeling problem (SPLP). Different to previous two-stage strategies that separate the detection and pose estimation steps, the SPLP model jointly infers the number of people, their poses, spatial proximity, and part level occlusions. Empirical results on four diverse and challenging datasets show significant improvements over all previous methods not only for the multi person, but also for the single person pose estimation problem. On multi person WAF dataset we improve by 30% PCP over the traditional two-stage approach. This shows that a joint formulation is crucial to disambiguate multiple and potentially overlapping persons. Models and code available at http://pose.mpi-inf.mpg.de.

References

[1] A. Alush and J. Goldberger. Ensemble segmentation using efficient integer linear programming. *TPAMI*, 34(10):1966–1977, 2012. 2

[2] B. Andres, J. H. Kappes, T. Beier, U. Köthe, and F. A. Hamprecht. Probabilistic image segmentation with closedness constraints. In *ICCV*, 2011. 2

[3] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR'14*. 1, 5, 6, 7, 8

[4] N. Bansal, A. Blum, and S. Chawla. Correlation clustering. *Machine Learning*, 56(1–3):89–113, 2004. 2

[5] V. Belagiannis, S. Amin, M. Andriluka, B. Schiele, N. Navab, and S. Ilic. 3D pictorial structures for multiple human pose estimation. In *CVPR'14*. 2

[6] 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. 5

[7] X. Chen and A. Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *NIPS'14*. 1, 2, 6, 8

[8] X. Chen and A. Yuille. Parsing occluded people by flexible compositions. In *CVPR*, 2015. 2, 7, 8

[9] S. Chopra and M. Rao. The partition problem. *Mathematical Programming*, 59(1–3):87–115, 1993. 2, 3

[10] E. D. Demaine, D. Emanuel, A. Fiat, and N. Immorlica. Correlation clustering in general weighted graphs. *Theoretical Computer Science*, 361(2–3):172–187, 2006. 2

[11] M. M. Deza and M. Laurent. *Geometry of Cuts and Metrics*. Springer, 1997. 2

[12] M. Eichner and V. Ferrari. We are family: Joint pose estimation of multiple persons. In *ECCV'10*. 2, 7

[13] G. Ghiasi, Y. Yang, D. Ramanan, and C. Fowlkes. Parsing occluded people. In *CVPR'14*. 7

[14] R. Girshick. Fast r-cnn. In *ICCV'15*. 4, 5

[15] G. Gkioxari, B. Hariharan, R. Girshick, and J. Malik. Using k-poselets for detecting people and localizing their keypoints. In *CVPR'14*. 1, 2

[16] H. Jiang and D. R. Martin. Global pose estimation using non-tree models. In *CVPR'09*. 2

[17] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC'10*. 5, 6

[18] S. Johnson and M. Everingham. Learning Effective Human Pose Estimation from Inaccurate Annotation. In *CVPR'11*. 1, 5

[19] S. Kim, C. Yoo, S. Nowozin, and P. Kohli. Image segmentation using higher-order correlation clustering. *TPAMI*, 36:1761–1774, 2014. 2

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS'12*. 5

[21] L. Ladicky, P. H. Torr, and A. Zisserman. Human pose estimation using a joint pixel-wise and part-wise formulation. In *CVPR'13*. 2

[22] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Strong appearance and expressive spatial models for human pose estimation. In *ICCV'13*. 2, 4

[23] L. Pishchulin, A. Jain, M. Andriluka, T. Thormaehlen, and B. Schiele. Articulated people detection and pose estimation: Reshaping the future. In *CVPR'12*. 1, 2

[24] D. Ramanan. Learning to parse images of articulated objects. In *NIPS'06*. 2

[25] X. Ren, A. C. Berg, and J. Malik. Recovering human body configurations using pairwise constraints between parts. In *ICCV'05*. 2

[26] B. Sapp and B. Taskar. Multimodal decomposable models for human pose estimation. In *CVPR'13*. 1, 2, 5

[27] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR,14*. 5

[28] M. Sun and S. Savarese. Articulated part-based model for joint object detection and pose estimation. In *ICCV'11*. 1, 2, 7

[29] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler. Efficient object localization using convolutional networks. In *CVPR'15*. 1, 2, 7

[30] J. 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'14*. 1, 5, 6, 7

[31] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR'14*. 2, 5

[32] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders. Selective search for object recognition. *IJCV'13*. 4

[33] X. Fan, K. Zheng, Y. Lin, and S. Wang. Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. In *CVPR'15*. 6

[34] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *PAMI'13*. 2, 7

[35] J. Yarkony, A. Ihler, and C. C. Fowlkes. Fast planar correlation clustering for image segmentation. In *ECCV*, 2012. 2