DeeperCut: A Deeper, Stronger, and Faster Multi-person Pose Estimation Model

Eldar Insafutdinov[1(✉)], Leonid Pishchulin[1], Bjoern Andres[1], Mykhaylo Andriluka[1,2], and Bernt Schiele[1]

[1] Max Planck Institute for Informatics, Saarbrücken, Germany
eldar@mpi-inf.mpq.de
[2] Stanford University, Stanford, USA

Abstract. The goal of this paper is to advance the state-of-the-art of articulated pose estimation in scenes with multiple people. To that end we contribute on three fronts. We propose (1) improved body part detectors that generate effective bottom-up proposals for body parts; (2) novel image-conditioned pairwise terms that allow to assemble the proposals into a variable number of consistent body part configurations; and (3) an incremental optimization strategy that explores the search space more efficiently thus leading both to better performance and significant speed-up factors. Evaluation is done on two single-person and two multi-person pose estimation benchmarks. The proposed approach significantly outperforms best known multi-person pose estimation results while demonstrating competitive performance on the task of single person pose estimation (Models and code available at http://pose.mpi-inf.mpg.de).

1 Introduction

Human pose estimation has recently made dramatic progress in particular on standard benchmarks for single person pose estimation [1,2]. This progress has been facilitated by the use of deep learning-based architectures [3,4] and by the availability of large-scale datasets such as "MPII Human Pose" [2]. In order to make further progress on the challenging task of multi-person pose estimation we carefully design and evaluate several key-ingredients for human pose estimation.

The first ingredient we consider is the generation of body part hypotheses. Essentially all prominent pose estimation methods include a component that detects body parts or estimates their position. While early work used classifiers such as SVMs and AdaBoost [1,5–7], modern approaches build on different flavors of deep learning-based architectures [8–11]. The second key ingredient are pairwise terms between body part hypotheses that help grouping those into valid human pose configurations. In earlier models such pairwise terms were essential for good performance [1,5,6]. Recent methods seem to profit less from such pairwise terms due to stronger unaries [8,10,11]. Image-conditioned pairwise terms [7,9] however have the promise to allow for better grouping. Last but not least, inference time is always a key consideration for pose estimation models. Often, model complexity has to be treated for speed and thus many

B. Leibe et al. (Eds.): ECCV 2016, Part VI, LNCS 9910, pp. 34–50, 2016.
DOI: 10.1007/978-3-319-46466-4_3

Fig. 1. Sample multi-person pose estimation results by the proposed *DeeperCut*.

models do not consider all spatial relations that would be beneficial for best performance.

In this paper we contribute to all three aspects and thereby significantly push the state of the art in multi-person pose estimation. We use a general optimization framework introduced in our previous work [10] as a test bed for all three key ingredients proposed in this paper, as it allows to easily replace and combine different components. Our contributions are three-fold, leading to a novel multi-person pose estimation approach that is deeper, stronger, and faster compared to the state of the art [10]:

- "deeper": we propose strong body part detectors based on recent advances in deep learning [12] that – taken alone – already allow to obtain competitive performance on pose estimation benchmarks.
- "stronger": we introduce novel image-conditioned pairwise terms between body parts that allow to push performance in the challenging case of multi-people pose estimation.
- "faster": we demonstrate that using our image-conditioned pairwise along with very good part detection candidates in a fully-connected model dramatically reduces the run-time by 2–3 orders of magnitude. Finally, we introduce a novel incremental optimization method to achieve a further 4x run-time reduction while improving human pose estimation accuracy.

We evaluate our approach on two single-person and two multi-person pose estimation benchmarks and report the best results in each case. Sample multi-person pose estimation predictions by the proposed approach are shown in Fig. 1.

Related work. Articulated human pose estimation has been traditionally formulated as a structured prediction task that requires an inference step combining local observations of body joints with spatial constraints. Various formulations have been proposed based on tree [6,13–15] and non-tree models [16,17]. The goal of the inference process has been to refine observations from local part detectors into coherent estimates of body configurations. Models of this type have been increasingly superseded by strong body part detectors [18–20], which has been reinforced by the development of strong image representations based on convolutional networks. Recent work aimed to incorporate convolutional detectors into part-based models [9] or design stronger detectors by combining the detector output with location-based features [21].

Specifically, as we suggest in [10], in the presence of strong detectors spatial reasoning results in diminishing returns because most contextual information can be incorporated directly in the detector. In this work we elevate the task to a new level of complexity by addressing images with multiple potentially overlapping people. This results in a more complex structured prediction problem with a variable number of outputs. In this setting we observe a large boost from conducting inference on top of state-of-the-art part detectors.

Combining spatial models with convnets allows to increase the receptive field that is used for inferring body joint locations. For example [11] iteratively trains a cascade of convolutional parts detectors, each detector taking the scoremap of all parts from the previous stage. This effectively increases the depth of the network and the receptive field is comparable to the entire person. With the recent developments in object detection newer architectures are composed of a large number of layers and the receptive field is large automatically. In this paper, we introduce a detector based on the recently proposed deep residual networks [12]. This allows us to train a detector with a large receptive field [11] and to incorporate intermediate supervision.

The use of purely geometric pairwise terms is suboptimal as they do not take local image evidence into account and only penalize deviation from the expected joint location. Due to the inherent articulation of body parts the expected location can only approximately guide the inference. While this can be sufficient when people are relatively distant from each other, for closely positioned people more discriminative pairwise costs are essential. Two prior works [7,9] have introduced image-dependent pairwise terms between connected body parts. While [7] uses an intermediate representation based on poselets our pairwise terms are conditioned directly on the image. [9] clusters relative positions of adjacent joints into $T = 11$ clusters, and assigns different labels to the part depending on which cluster it falls to. Subsequently a CNN is trained to predict this extended set of classes and later an SVM is used to select the maximum scoring joint pair relation.

Single person pose estimation has advanced considerably, but the setting is simplified. Here we focus on the more challenging problem of multi-person pose estimation. Previous work has addressed this problem as sequence of person detection and pose estimation [22–24]. [22] use a detector for initialization and

reasoning across people, but rely on simple geometric body part relationships and only reason about person-person occlusions. [24] focus on single partially occluded people, and handle multi-person scenes akin to [6]. In [10] we propose to jointly detect and estimate configurations, but rely on simple pairwise terms only, which limits the performance and, as we show, results in prohibitive inference time to fully explore the search space. Here, we innovate on multiple fronts both in terms of speed and accuracy.

2 DeepCut Recap

This section summarizes *DeepCut* [10] and how unary and pairwise terms are used in this approach. *DeepCut* is a state-of-the-art approach to multi-person pose estimation based on integer linear programming (ILP) that jointly estimates poses of all people present in an image by minimizing a joint objective. This objective aims to jointly partition and label an initial pool of body part candidates into consistent sets of body-part configurations corresponding to distinct people. We use *DeepCut* as a general optimization framework that allows to easily replace and combine different components.

Specifically, *DeepCut* starts from a set D of *body part candidates*, i.e. putative detections of body parts in a given image, and a set C of *body part classes*, e.g., head, shoulder, knee. The set D of part candidates is typically generated by body part detectors and each candidate $d \in D$ has a *unary score* for every body part class $c \in C$. Based on these unary scores *DeepCut* associates a cost or reward $\alpha_{dc} \in \mathbb{R}$ to be paid by all feasible solutions of the pose estimation problem for which the body part candidate d is a body part of class c.

Additionally, for every pair of distinct body part candidates $d, d' \in D$ and every two body part classes $c, c' \in C$, the *pairwise term* is used to generate a cost or reward $\beta_{dd'cc'} \in \mathbb{R}$ to be paid by all feasible solutions of the pose estimation problem for which the body part d, classified as c, and the body part d', classified as c', belong to the same person.

With respect to these sets and costs, the pose estimation problem is cast as an ILP in two classes of 01-variables: Variables $x : D \times C \to \{0, 1\}$ indicate by $x_{dc} = 1$ that body part candidate d is of body part class c. If, for a $d \in D$ and all $c \in C$, $x_{dc} = 0$, the body part candidate d is suppressed. Variables $y : \binom{D}{2} \to \{0, 1\}$ indicate by $y_{dd'} = 1$ that body part candidates d and d' belong to the same person. Additional variables and constraints described in [10] link the variables x and y to the costs and ensure that feasible solutions (x, y) well-define a selection and classification of body part candidates as body part classes as well as a clustering of body part candidates into distinct people.

The *DeepCut* ILP is hard and hard to approximate, as it generalizes the minimum cost multicut or correlation clustering problem which is APX-hard [25,26]. Using the branch-and-cut algorithm [10] to compute constant-factor approximative feasible solutions of instances of the *DeepCut* ILP is not necessarily practical. In Sect. 5 we propose an incremental optimization approach that uses branch-and-cut algorithm to incrementally solve several instances of ILP, which results into 4–5x run-time reduction with increased pose estimation accuracy.

3 Part Detectors

As argued before, strong part detectors are an essential ingredient of modern pose estimation methods. We propose and evaluate a deep fully-convolutional human body part detection model drawing on powerful recent ideas from semantic segmentation, object classification [12,27,28] and human pose estimation [10,11,20].

3.1 Model

Architecture. We build on the recent advances in object classification and adapt the extremely deep Residual Network (ResNet) [12] for human body part detection. This model achieved excellent results on the recent ImageNet Object Classification Challenge and specifically tackles the problem of vanishing gradients by passing the state through identity layers and modeling residual functions. Our best performing body part detection model has 152 layers (c.f. Sect. 3.2) which is in line with the findings of [12].

Stride. Adapting ResNet for the sliding window-based body part detection is not straight forward: converting ResNet to the fully convolutional mode leads to a 32 px stride which is too coarse for precise part localization. In [10] we show that using a stride of 8 px leads to good part detection results. Typically, spatial resolution can be recovered by either introducing up-sampling *deconvolutional* layers [27], or blowing up the convolutional filters using the *hole algorithm* [28]. The latter has shown to perform better on the task of semantic segmentation. However, using the *hole algorithm* to recover the spatial resolution of ResNet is infeasible due to memory constraints. For instance, the 22 residual blocks in the conv4 bank of ResNet-101 constitute the major part of the network and running it at stride 8 px does not fit the net into GPU memory[1]. We thus employ a hybrid approach. First, we remove the final classification as well as average pooling layer. Then, we decrease the stride of the first convolutional layers of the conv5 bank from 2 px to 1 px to prevent down-sampling. Next, we add holes to all 3x3 convolutions in conv5 to preserve their receptive field. This reduces the stride of the full CNN to 16 px. Finally, we add deconvolutional layers for 2x up-sampling and connect the final output to the output of the conv3 bank.

Receptive field size. A large receptive field size allows to incorporate context when predicting locations of individual body parts. [8,11] argue about the importance of large receptive fields and propose a complex hierarchical architecture predicting parts at multiple resolution levels. The extreme depth of ResNet allows for a very large receptive field (on the order of 1000 px compared to VGG's 400 px [4]) without the need of introducing complex hierarchical architectures. We empirically find that re-scaling the original image such that an upright standing person is 340 px high leads to best performance.

Intermediate supervision. Providing additional supervision addresses the problem of vanishing gradients in deep neural networks [11,29,30]. In addition

[1] We use NVIDIA Tesla K40 GPU with 12 GB RAM.

to that, [11] reports that using part scoremaps produced at intermediate stages as inputs for subsequent stages helps to encode spatial relations between parts, while [31] use spatial fusion layers that learn an implicit spatial model. ResNets address the first problem by introducing identity connections and learning residual functions. To address the second concern, we make a slightly different choice: we add part loss layers inside the conv4 bank of ResNet. We argue that it is not strictly necessary to use scoremaps as inputs for the subsequent stages. The activations from such intermediate predictions are different only up to a linear transformation and contain all information about part presence that is available at that stage of the network. In Sect. 3.2 we empirically show a consistent improvement of part detection performance when including intermediate supervision.

Loss functions. We use sigmoid activations and cross entropy loss function during training [10]. We perform location refinement by predicting offsets from the locations on the scoremap grid to the ground truth joint locations [10].

Training. We use the publicly available ResNet implementation (Caffe) and initialize from the ImageNet-pre-trained models. We train networks with SGD for 1M iterations, starting with the learning rate lr=0.001 for 10k, then lr=0.002 for 420k, lr=0.0002 for 300k and lr=0.0001 for 300k. This corresponds to roughly 17 epochs of the MPII [2] train set. Finetuning from ImageNet takes two days on a *single* GPU. Batch normalization [32] worsens performance, as the batch size of 1 in fully convolutional training is not enough to provide a reliable estimate of activation statistics. During training we switch off collection of statistics and use the mean and variance that were gathered on the ImageNet dataset.

3.2 Evaluation of Part Detectors

Datasets. We use three public datasets: "Leeds Sports Poses" (LSP) [1] (person-centric (PC) annotations); "LSP Extended" (LSPET) [15]; "MPII Human Pose" ("Single Person") [2] consisting of 19185 training and 7247 testing poses. To evaluate on LSP we train part detectors on the union of MPII, LSPET and LSP training sets. To evaluate on MPII Single Person we train on MPII *only*.

Evaluation measures. We use the standard "Percentage of Correct Keypoints (PCK)" evaluation metric [8,33,34] and evaluation scripts from the web page of [2]. In addition to PCK at fixed threshold, we report "Area under Curve" (AUC) computed for the entire range of PCK thresholds.

Results on LSP. The results are shown in Table 1. ResNet-50 with 8 px stride achieves 87.8 % PCK and 63.7 % AUC. Increasing the stride size to 16 px and up-sampling the scoremaps by 2x to compensate for the loss on resolution slightly drops the performance to 87.2 % PCK. This is expected as up-sampling cannot fully compensate for the information loss due to a larger stride. Larger stride minimizes memory requirements, which allows for training a deeper ResNet-152. The latter significantly increases the performance (89.1 vs. 87.2 % PCK, 65.1 vs. 63.1 % AUC), as it has larger model capacity. Introducing intermediate supervision further improves the performance to 90.1 % PCK and 66.1 % AUC, as it constraints

Table 1. Pose estimation results (PCK) on LSP (PC) dataset.

Setting	Head	Sho	Elb	Wri	Hip	Knee	Ank	PCK	AUC
ResNet-50 (8 px)	96.9	90.3	85.0	81.5	88.6	87.3	84.8	87.8	63.7
ResNet-50 (16 px + 2x up-sample)	96.7	89.8	84.6	80.4	89.3	86.4	82.8	87.2	63.1
ResNet-101 (16 px + 2x up-sample)	96.9	91.2	85.8	82.6	90.9	**90.2**	85.9	89.1	64.6
ResNet-152 (16 px + 2x up-sample)	97.4	91.7	85.7	82.4	90.1	89.2	86.9	89.1	65.1
+ intermediate supervision	97.4	**92.7**	**87.5**	**84.4**	**91.5**	89.9	87.2	90.1	**66.1**
DeepCut [10]	97.0	91.0	83.8	78.1	91.0	86.7	82.0	87.1	63.5
Wei et al. [11]	**97.8**	92.5	87.0	83.9	**91.5**	90.8	**89.9**	**90.5**	65.4
Tompson et al. [8]	90.6	79.2	67.9	63.4	69.5	71.0	64.2	72.3	47.3
Chen & Yuille [9]	91.8	78.2	71.8	65.5	73.3	70.2	63.4	73.4	40.1
Fan et al. [35]	92.4	75.2	65.3	64.0	75.7	68.3	70.4	73.0	43.2

the network to learn useful representations in the early stages and uses them in later stages for spatial disambiguation of parts.

The results are compared to the state of the art in Table 1. Our best model significantly outperforms *DeepCut* [10] (90.1 % PCK vs. 87.1 % PCK), as it relies on deeper detection architectures. Our model performs on par with the recent approach of Wei et al. [11] (90.1 vs. 90.5 % PCK, 66.1 vs. 65.4 AUC). This is interesting, as they use a much more complex multi-scale multi-stage architecture.

Results on MPII Single Person. The results are shown in Table 2. ResNet-152 achieves 87.8 % PCK_h and 60.0 % AUC, while intermediate supervision slightly improves the performance further to 88.5 % PCK_h and 60.8 % AUC. Comparing the results to the state of the art we observe significant improvement over *DeepCut* [10] (+5.9 % PCK_h, +4.2 % AUC), which again underlines the

Table 2. Pose estimation results (PCK_h) on MPII Single Person.

Setting	Head	Sho	Elb	Wri	Hip	Knee	Ank	PCK_h	AUC
ResNet-152	96.3	94.1	88.6	83.9	87.2	82.9	77.8	87.8	60.0
+ intermediate supervision	96.8	**95.2**	**89.3**	**84.4**	**88.4**	**83.4**	78.0	**88.5**	60.8
DeepCut [10]	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4	56.5
Tompson et al. [8]	95.8	90.3	80.5	74.3	77.6	69.7	62.8	79.6	51.8
Carreira et al. [36]	95.7	91.7	81.7	72.4	82.8	73.2	66.4	81.3	49.1
Tompson et al. [20]	96.1	91.9	83.9	77.8	80.9	72.3	64.8	82.0	54.9
Wei et al. [11]	**97.8**	95.0	88.7	84.0	**88.4**	82.8	**79.4**	**88.5**	**61.4**

importance of using extremely deep model. The proposed approach performs on par with the best know result by Wei et al. [11] (88.5 vs. 88.5 % PCK_h) for the maximum distance threshold, while slightly loosing when using the entire range of thresholds (60.8 vs. 61.4 % AUC). We envision that extending the proposed approach to incorporate multiple scales as in [11] should improve the performance. The model trained on the union of MPII, LSPET and LSP training sets achieves 88.3 % PCK_h and 60.7 % AUC. The fact that we use the same trained model on both LSP and MPII benchmarks and achieve similar performance demonstrates the generality of the proposed approach.

4 Image-Conditioned Pairwise Terms

As discussed in Sect. 3, a large receptive field for the CNN-based part detectors allows to accurately predict the presence of a body part at a given location. However, it also contains enough evidence to reason about locations of other parts in the vicinity. We draw on this insight and propose to also use deep networks to make pairwise part-to-part predictions. They are subsequently used to compute the pairwise probabilities and show significant improvements for multi-person pose estimation.

4.1 Model

Our approach is inspired by the body part location refinement described in Sect. 3. In addition to predicting offsets for the current joint, we directly regress from the current location to the relative positions of all other joints. For each scoremap location $k = (x_k, y_k)$ that is marked positive w.r.t the joint $c \in C$ and for each remaining joint $c' \in C \setminus c$, we define a relative position of c' w.r.t. c as a tuple $t^k_{cc'} = (x_{c'} - x_k, y_{c'} - x_k)$. We add an extra layer that predicts relative position $o^k_{cc'}$ and train it with a smooth L_1 loss function. We thus perform *joint* training of body part detectors (cross-entropy loss), location regression (L_1 loss) and pairwise regression (L_1 loss) by linearly combining all three loss functions. The targets t are normalized to have zero mean and unit variance over the training set. Results of such predictions are shown in Fig. 2.

We then use these predictions to compute pairwise costs $\beta_{dd'cc'}$. For any pair of detections (d, d') (Fig. 3) and for any pair of joints (c, c') we define the following quantities: locations l_d, l'_d of detections d and d' respectively; the offset prediction $o^d_{cc'}$ from c to c' at location d (solid red) coming from the CNN and similarly the offset prediction $o^{d'}_{c'c}$ (solid turquoise). We then compute the offset between the two predictions: $\hat{o}_{dd'} = l_{d'} - l_d$ (marked in dashed red). The degree to which the prediction $o^d_{cc'}$ agrees with the actual offset $\hat{o}_{dd'}$ tells how likely d, d' are of classes c, c' respectively and belong to the same person. We measure this by computing the distance between the two offsets $\Delta_f = \|\hat{o}_{dd'} - o^d_{cc'}\|_2$, and the absolute angle $\theta_f = |\measuredangle(\hat{o}_{dd'}, o^d_{cc'})|$ where f stands for forward direction, i.e. from d to d'. Similarly, we incorporate the prediction $o^{d'}_{c'c}$ in the backwards direction by computing $\Delta_b = \|\hat{o}_{d'd} - o^{d'}_{c'c}\|_2$ and $\theta_b = |\measuredangle(\hat{o}_{d'd}, o^{d'}_{c'c})|$. Finally, we

Fig. 2. Visualizations of regression predictions. Top: from left shoulder to the right shoulder (green), right hip (red), left elbow (light blue), right ankle (purple) and top of the head (dark blue). Bottom: from right knee to the right hip (green), right ankle (red), left knee (dark blue), left ankle (light blue) and top of the head (purple). Longer-range predictions, such as e.g. shoulder – ankle may be less accurate for harder poses (top row, images 2 and 3) compared to the nearby predictions. However, they provide enough information to constrain the search space in the fully-connected spatial model. (Color figure online)

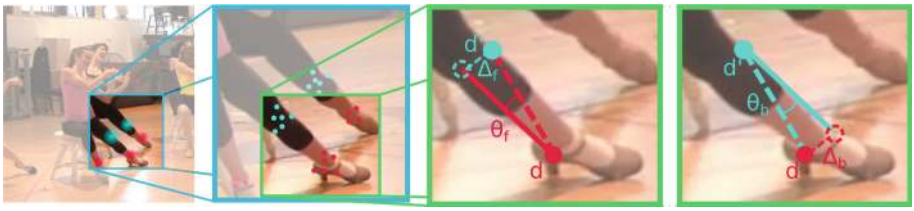

Fig. 3. Visualization of features extracted to score the pairwise. See text for details. (Color figure online)

define a feature vector by augmenting features with exponential terms: $f_{dd'cc'} = (\Delta_f, \theta_f, \Delta_b, \theta_b, \exp(-\Delta_f), \ldots, \exp(-\theta_b))$.

We then use the features $f_{dd'cc'}$ and define logistic model:

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

where $K = (|C| \times (|C| + 1))/2$ parameters $\omega_{cc'}$ are estimated using ML.

4.2 Sampling Detections

Location refinement NMS. *DeepCut* samples the set of detections D from the scoremap by applying non-maximum suppression (NMS). Here, we utilize location refinement and correct grid locations with the predicted offsets before applying NMS. This pulls detections that belong to a particular body joint

Table 3. Effects of proposed pairwise and unaries on the pose estimation performance (AP) on MPII Multi-person Val.

Unary	Pairwise	Head	Sho	Elb	Wri	Hip	Knee	Ank	AP	time [s/frame]
DeepCut [10]	*DeepCut* [10]	50.1	44.1	33.5	26.5	33.0	28.5	14.4	33.3	259220
DeepCut [10]	this work	68.3	58.3	47.4	38.9	45.2	41.8	31.2	47.7	1987
this work	this work	**70.9**	59.8	**53.1**	**44.4**	50.0	46.4	39.5	52.3	1171
+ location refinement before NMS		70.3	**61.6**	52.1	43.7	**50.6**	**47.0**	**40.6**	**52.6**	**578**

towards its true location thereby increasing the density of detections around that location, which allows to distribute the detection candidates in a better way.

Splitting of part detections. *DeepCut* ILP solves the clustering problem by labeling each detection d with a single part class c and assigning it to a particular cluster that corresponds to a distinct person. However, it may happen that the same spatial location is occupied by more than one body joint, and therefore, its corresponding detection can only be labeled with one of the respecting classes. A naive solution is to replace a detection with n detections for each part class, which would result in a prohibitive increase in the number of detections. We simply split a detection d into several if more than one part has unary probability that is higher than a chosen threshold s (in our case $s = 0.4$).

4.3 Evaluation of Pairwise

Datasets and evaluation measure. We evaluate on the challenging public "MPII Human Pose" ("Multi-Person") benchmark [2] consisting of 3844 training and 1758 testing groups of multiple overlapping people in highly articulated poses with a variable number of parts. We perform all intermediate experiments on a validation set of 200 images sampled uniformly at random and refer to it as MPII Multi-Person Val. We report major results on the full testing set, and on the subset of 288 images for the direct comparison to [10]. The AP measure [10] evaluating consistent body part detections is used for performance comparison. Additionally, we report median running time per frame measured in seconds[2].

Table 4. Effects of different versions of the pairwise terms on the pose estimation performance (AP) on MPII Multi-person Val.

Setting	Head	Sho	Elb	Wri	Hip	Knee	Ank	AP	time [s/frame]
bi-directional + angle	**70.3**	**61.6**	**52.1**	43.7	**50.6**	**47.0**	**40.6**	**52.6**	**578**
uni-directional + angle	69.3	58.4	51.8	**44.2**	50.4	44.7	36.3	51.1	2140
bi-directional	68.8	58.3	51.0	42.7	51.1	46.5	38.7	51.3	914

Evaluation of unaries and pairwise. The results are shown in Table 3. Baseline *DeepCut* achieves 33.3 % AP. Using the proposed pairwise significantly

[2] Run-time is measured on a single core Intel Xeon 2.70 GHz.

improves performance achieving 47.7 % AP. This clearly shows the advantages of using image-conditioned pairwise to disambiguate the body part assignment for multiple overlapping individuals. Remarkably, the proposed pairwise dramatically reduce the run-time by two orders of magnitude (1987 vs. 259220 s/frame). This underlines the argument that using strong pairwise in the fully-connected model allows to significantly speed-up the inference. Using additionally the proposed part detectors further boosts the performance (52.3 vs. 47.7 % AP), which can be attributed to better quality part hypotheses. Run-time is again almost halved, which clearly shows the importance of obtaining high-quality part detection candidates for more accurate and faster inference. Performing location refinement before NMS slightly improves the performance, but also reduces the run-time by 2x: this allows to increase the density of detections at the most probable body part locations and thus suppresses more detections around the most confident ones, which leads to better distribution of part detection candidates and reduces confusion generated by the near-by detections. Overall, we observe significant performance improvement and dramatic reduction in run-time by the proposed *DeeperCut* compared to the baseline *DeepCut*.

Ablation study of pairwise. An ablation study of the proposed image-conditioned pairwise is performed in Table 4. Regressing from both joints onto the opposite joint's location and including angles achieves the best performance of 52.6 % AP and the minimum run-time of 578 s/frame. Regressing from a single joint only slightly reduces the performance to 51.1 % AP, but significantly increases run-time by 4x: these pairwise are less robust compared to the bi-directional, which confuses the inference. Removing the angles from the pairwise features also decreases the performance (51.3 vs. 52.6 % AP) and doubles run-time, as it removes the information about body part orientation.

5 Incremental Optimization

Solving one instance of the *DeepCut* ILP for all body part candidates detected for an image, as suggested in [10] and summarized in Sect. 2, is elegant in theory but disadvantageous in practice:

Firstly, the time it takes to compute constant-factor approximative feasible solution by the branch-and-cut algorithm [10] can be exponential in the number of body part candidates in the worst case. In practice, this limits the number of candidates that can be processed by this algorithm. Due to this limitation, it does happen that body parts and, for images showing many persons, entire persons are missed, simply because they are not contained in the set of candidates.

Secondly, solving one instance of the optimization problem for the entire image means that no distinction is made between part classes detected reliably, e.g. head and shoulders, and part classes detected less reliably, e.g. wrists, elbows and ankles. Therefore, it happens that unreliable detections corrupt the solution.

To address both problems, we solve not one instance of the *DeepCut* ILP but several, starting with only those body part classes that are detected most reliably

and only then considering body part classes that are detected less reliably. Concretely, we study two variants of this incremental optimization approach which are defined in Table 5. Specifically, the procedure works as follows:

For each subset of body part classes defined in Table 5, an instance of the *DeepCut* ILP is set up and a constant-factor approximative feasible solution computed using the branch-and-cut algorithm. This feasible solution selects, labels and clusters a subset of part candidates, namely of those part classes that are considered in this instance. For the next instance, each cluster of body part candidates of the same class from the previous instance becomes just one part candidate whose class is fixed. Thus, the next instance is an optimization problem for selecting, labeling and clustering body parts that have not been determined by previous instances. Overall, this allows us to start with more part candidates consistently and thus improve the pose estimation result significantly.

Table 5. As the run-time of the DeepCut branch-and-cut algorithm limits the number of part candidates that can be processed in practice, we split the set of part classes into subsets, coarsely and finely, and solve the pose estimation problem incrementally.

	Stage 1	Stage 2	Stage 3
2-stage	head, shoulders elbows, wrists	hips, knees ankles	
3-stage	head shoulders	elbows wrists	hips, knees ankles

5.1 Evaluation of Incremental Optimization

Results are shown in Table 6. Single stage optimization with $|D| = 100$ part detection candidates achieves 52.6 % AP (best from Table 3). More aggressive NMS with radius of 24 px improves the performance (54.5 vs. 52.6 % AP), as it allows to better distribute detection candidates. Increasing $|D|$ to 150 slightly improves the performance by +0.6 % AP, but significantly increases run-time (1041 vs. 596 s/frame). We found $|D| = 150$ to be maximum total number of detection candidates (11 per part) for which optimization runs in a reasonable time. Incremental optimization of 2-stage inference slightly improves the performance (56.5 vs. 55.1 % AP) as it allows for a larger number of detection candidates per body part (20) and leverages typically more confident predictions of the upper body parts in the first stage before solving for the entire body. Most importantly, it halves the median run-time from 1041 to 483 s/frame. Incremental optimization of 3-stage inference again almost halves the run-time to 271 s/frame while noticeably improving the human pose estimation performance for all body parts but elbows achieving 57.6 % AP. These results clearly demonstrate the advantages of the proposed incremental optimization. Splitting the detection candidates that simultaneously belong to multiple body parts with

Table 6. Performance (AP) of different hierarchical versions of *DeeperCut* on MPII Multi-person Val.

Setting	Head	Sho	Elb	Wri	Hip	Knee	Ank	AP	time [s/frame]
1-stage optimize, 100 det, nms 1x	70.3	61.6	52.1	43.7	50.6	47.0	40.6	52.6	578
1-stage optimize, 100 det, nms 2x	71.3	64.1	55.8	44.1	53.8	48.7	41.3	54.5	596
1-stage optimize, 150 det, nms 2x	74.1	65.6	56.0	44.3	54.4	49.2	39.8	55.1	1041
2-stage optimize	75.9	66.8	58.8	46.1	54.1	48.7	42.4	56.5	483
3-stage optimize	78.3	69.3	58.4	47.5	55.1	49.6	42.5	57.6	271
+ split detections	**78.5**	**70.5**	**59.7**	**48.7**	**55.4**	**50.6**	**44.4**	**58.7**	**270**
DeepCut [10]	50.1	44.1	33.5	26.5	33.0	28.5	14.4	33.3	259220

high confidence slightly improves the performance to 58.7 % AP. This helps to overcome the limitation that each detection candidate can be assigned to a single body part and improves on cases where two body parts overlap thus sharing the same detection candidate. We also compare the obtained results to *DeepCut* in Table 6 (last row). The proposed *DeeperCut* outperforms baseline *DeepCut* (58.7 vs. 33.3 % AP) by almost doubling the performance, while run-time is reduced dramatically by 3 orders of magnitude from the infeasible 259220 s/frame to affordable 270 s/frame. This comparison clearly demonstrates the power of the proposed approach and dramatic effects of better unary, pairwise and optimization on the overall pose estimation performance and run-time.

5.2 Comparison to the State of the Art

We compare to others on MPII Multi-Person Test and WAF [22] datasets.

Results on MPII Multi-person. For direct comparison with *DeepCut* we evaluate on the same subset of 288 testing images as in [10]. Additionally, we provide the results on the entire testing set. Results are shown in Table 7. *DeeperCut* without incremental optimization already outperforms *DeepCut* by a large margin (66.2 vs. 54.1 % AP). Using 3-stage incremental optimization further improves the performance to 69.7 % AP improving by a dramatic 16.5 % AP over the baseline. Remarkably, the run-time is reduced from 57995 to 230 s/frame, which is an improvement by two orders of magnitude. Both results underline the importance of strong image-conditioned pairwise terms and incremental optimization to maximize multi-person pose estimation performance at the reduced run-time. A similar trend is observed on the full set: 3-stage optimization improves over a single stage optimization (59.4 vs. 54.7 % AP). We observe that the performance on the entire testing set is over 10 % AP lower compared to the subset and run-time is doubled. This implies that the subset of 288 images is easier compared to the full testing set. We envision that performance differences between *DeeperCut* and *DeepCut* on the entire set will be

Table 7. Pose estimation results (AP) on MPII Multi-person.

Setting	Head	Sho	Elb	Wri	Hip	Knee	Ank	AP	time [s/frame]
subset of 288 images as in [10]									
DeeperCut (1-stage)	83.3	79.4	66.1	57.9	63.5	60.5	49.9	66.2	1333
DeeperCut	**87.5**	**82.8**	**70.2**	**61.6**	**66.0**	**60.6**	**56.5**	**69.7**	**230**
DeepCut [10]	73.4	71.8	57.9	39.9	56.7	44.0	32.0	54.1	57995
full set									
DeeperCut (1-stage)	73.7	65.4	54.9	45.2	52.3	47.8	40.7	54.7	2785
DeeperCut	**79.1**	**72.2**	**59.7**	**50.0**	**56.0**	**51.0**	**44.6**	**59.4**	485
Faster R-CNN [37] + unary	64.9	62.9	53.4	44.1	50.7	43.1	35.2	51.0	**1**

at least as large as when compared on the subset. We also compare to a strong two-stage baseline: first each person is pre-localized by applying the state-of-the-art detector [37] following by NMS and retaining rectangles with scores at least 0.8; then pose estimation for each rectangle is performed using *DeeperCut* unary only. Being significantly faster (1 s/frame) this approach reaches 51.0 % AP vs. 59.4 % AP by *DeeperCut*, which clearly shows the power of joint reasoning by the proposed approach.

Table 8. Pose estimation results (mPCP) on WAF dataset.

Setting	Head	U Arms	L Arms	Torso	mPCP	AOP
DeeperCut nms 3.0	**99.3**	**83.8**	**81.9**	**87.1**	**86.3**	**88.1**
DeepCut [10]	**99.3**	81.5	79.5	**87.1**	84.7	86.5
Ghiasi et al. [38]	-	-	-	-	63.6	74.0
Eichner & Ferrari [22]	97.6	68.2	48.1	86.1	69.4	80.0
Chen & Yuille [24]	98.5	77.2	71.3	88.5	80.7	84.9

Results on WAF. Results using the official evaluation protocol [22] assuming mPCP and AOP evaluation measures and considering detection bounding boxes provided by [22] are shown in Table 8. *DeeperCut* achieves the best result improving over the state of the art *DeepCut* (86.3 vs. 84.7 % mPCP, 88.1 vs. 86.5 % AOP). Noticeable improvements are observed both for upper (+2.3 % mPCP) and lower (+2.4 % mPCP) arms. However, overall performance differences between *DeeperCut* and the baseline *DeepCut* are not as pronounced compared to MPII Multi-Person dataset. This is due to the fact that actual differences are washed out by the peculiarities of the mPCP evaluation measure: mPCP assumes that people are pre-detected and human pose estimation performance is evaluated only for people whose upper body detections match the ground truth. Thus, a pose estimation method is not penalized for generating multiple body pose predictions, since the only pose prediction is considered

whose upper body bounding box best matches the ground truth. We thus re-evaluate the competing approaches [10,24] using the more realistic AP evaluation measure[3]. The results are shown in Table 9. *DeeperCut* significantly improves over *DeepCut* (82.0 vs. 76.2 % AP). The largest boost in performance is achieved for head (+16.0 % AP) and wrists (+5.2 % AP): *DeeperCut* follows incremental optimization strategy by first solving for the most reliable body parts, such as head and shoulders, and then using the obtained solution to improve estimation of harder body parts, such as wrists. Most notably, run-time is dramatically reduced by 3 orders of magnitude from 22000 to 13 s/frame. These results clearly show the advantages of the proposed approach when evaluated in the real-world detection setting. The proposed *DeeperCut* also outperforms [24] by a large margin. The performance difference is much more pronounced compared to using mPCP evaluation measure: in contrast to mPCP, AP penalizes multiple body pose predictions of the same person. We envision that better NMS strategies are likely to improve the AP performance of [24].

Table 9. Pose estimation results (AP) on WAF dataset.

Setting	Head	Sho	Elb	Wri	AP	time [s/frame]
DeeperCut	**92.6**	**81.1**	**75.7**	**78.8**	**82.0**	**13**
DeepCut [10]	76.6	80.8	73.7	73.6	76.2	22000
Chen & Yuille [24]	83.3	56.1	46.3	35.5	55.3	-

6 Conclusion

In this paper we significantly advanced the state of the art in articulated multi-person human pose estimation. To that end we carefully re-designed and thoroughly evaluated several key ingredients. First, drawing on the recent advances in deep learning we proposed strong extremely deep body part detectors that – taken alone – already allow to obtain state of the art performance on standard pose estimation benchmarks. Second, we introduce novel image-conditioned pairwise terms between body parts that allow to significantly push the performance in the challenging case of multi-people pose estimation, and dramatically reduce the run-time of the inference in the fully-connected spatial model. Third, we introduced a novel incremental optimization strategy to further reduce the run-time and improve human pose estimation accuracy. Overall, the proposed improvements allowed to almost double the pose estimation accuracy in the challenging multi-person case while reducing the run-time by 3 orders of magnitude.

[3] We used publicly-available pose predictions of [24] for all people in WAF dataset.

References

1. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. In: BMVC 2010
2. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: CVPR 2014
3. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS 2012
4. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: CoRR 2014
5. Andriluka, M., Roth, S., Schiele, B.: Discriminative appearance models for pictorial structures. In: IJCV 2011
6. Yang, Y., Ramanan, D.: Articulated human detection with flexible mixtures of parts. In: PAMI 2013
7. Pishchulin, L., Andriluka, M., Gehler, P., Schiele, B.: Poselet conditioned pictorial structures. In: CVPR 2013
8. Tompson, J.J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. In: NIPS 2014
9. Chen, X., Yuille, A.: Articulated pose estimation by a graphical model with image dependent pairwise relations. In: NIPS 2014
10. Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P., Schiele, B.: Deepcut: Joint subset partition and labeling for multi person pose estimation. In: CVPR 2016
11. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: CVPR 2016
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR 2016
13. Ramanan, D.: Learning to parse images of articulated objects. In: NIPS 2006
14. Jiang, H., Martin, D.R.: Global pose estimation using non-tree models. In: CVPR 2009
15. Johnson, S., Everingham, M.: Learning effective human pose estimation from inaccurate annotation. In: CVPR 2011
16. Tran, D., Forsyth, D.: Improved human parsing with a full relational model. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 227–240. Springer, Heidelberg (2010). doi:10.1007/978-3-642-15561-1_17
17. Wang, F., Li, Y.: Beyond physical connections: Tree models in human pose estimation. In: CVPR 2013
18. Pishchulin, L., Andriluka, M., Gehler, P., Schiele, B.: Strong appearance and expressive spatial models for human pose estimation. In: ICCV 2013
19. Gkioxari, G., Arbelaez, P., Bourdev, L., Malik, J.: Articulated pose estimation using discriminative armlet classifiers. In: CVPR 2013
20. Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C.: Efficient object localization using convolutional networks. In: CVPR 2015
21. Ramakrishna, V., Munoz, D., Hebert, M., Andrew Bagnell, J., Sheikh, Y.: Pose machines: Articulated pose estimation via inference machines. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part II. LNCS, vol. 8690, pp. 33–47. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10605-2_3
22. Eichner, M., Ferrari, V.: We are family: Joint pose estimation of multiple persons. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 228–242. Springer, Heidelberg (2010). doi:10.1007/978-3-642-15549-9_17

23. Ladicky, L., Torr, P.H., Zisserman, A.: Human pose estimation using a joint pixel-wise and part-wise formulation. In: CVPR 2013
24. Chen, X., Yuille, A.: Parsing occluded people by flexible compositions. In: CVPR 2015
25. Bansal, N., Blum, A., Chawla, S.: Correlation clustering. In: ML 2004
26. Demaine, E.D., Emanuel, D., Fiat, A., Immorlica, N.: Correlation clustering in general weighted graphs. In: Theoretical Computer Science 2006
27. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR 2015
28. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. In: ICLR 2015
29. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR 2015
30. Lee, C.Y., Xie, S., Gallagher, P., Zhang, Z., Tu, Z.: Deeply-supervised nets. In: AISTATS 2015
31. Pfister, T., Charles, J., Zisserman, A.: Flowing convnets for human pose estimation in videos. In: ICCV 2015
32. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: CoRR 2015
33. Sapp, B., Taskar, B.: Multimodal decomposable models for human pose estimation. In: CVPR 2013
34. Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In: CVPR 2014
35. Fan, X., Zheng, K., Lin, Y., Wang, S.: Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. In: CVPR 2015
36. Carreira, J., Agrawal, P., Fragkiadaki, K., Malik, J.: Human pose estimation with iterative error feedback. In: CVPR 2016
37. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: NIPS 2015
38. Ghiasi, G., Yang, Y., Ramanan, D., Fowlkes, C.: Parsing occluded people. In: CVPR 2014