Deep Fully-Connected Part-Based Models for Human Pose Estimation

Rodrigo de Bem RODRIGO@ROBOTS.OX.AC.UK
University of Oxford, UK and Federal University of Rio Grande, Brazil
Anurag Arnab AARNAB@ROBOTS.OX.AC.UK
University of Oxford, UK
Stuart Golodetz STUART.GOLODETZ@ENG.OX.AC.UK
University of Oxford, UK
Michael Sapienza M.SAPIENZA@SAMSUNG.COM
Think Tank Team, Samsung Research America, Mountain View, USA
Philip Torr PHILIP.TORR@ENG.OX.AC.UK
University of Oxford, UK

Editors: Jun Zhu and Ichiro Takeuchi

Abstract

We propose a 2D multi-level appearance representation of the human body in RGB images, spatially modelled using a fully-connected graphical model. The appearance model is based on a CNN body part detector, which uses shared features in a cascade architecture to simultaneously detect body parts with different levels of granularity. We use a fully-connected Conditional Random Field (CRF) as our spatial model, over which approximate inference is efficiently performed using the Mean-Field algorithm, implemented as a Recurrent Neural Network (RNN). The stronger visual support from body parts with different levels of granularity, along with the fully-connected pairwise spatial relations, which have their weights learnt by the model, improve the performance of the bottom-up part detector. We adopt an end-to-end training strategy to leverage the potential of both our appearance and spatial models, and achieve competitive results on the MPII and LSP datasets.

Keywords: Deep learning, fully-connected part-based models, CRFs, human pose estimation

1. Introduction

Human pose estimation is a fundamental problem in computer vision, with important applications in human-computer interaction, motion capture for films and games, activity recognition and prediction and augmented reality (Moeslund et al., 2011). Unconstrained monocular pose estimation presents several challenges. The human body, modelled as an articulated object, lies in a high-dimensional space in which finding feasible configurations is costly. The projection from the 3D world to the 2D image plane is a source of uncertainties and ambiguities that must also be overcome. Lastly, one can mention the huge variety of sizes, shapes, appearances and poses that the human body can assume, and the endless number of different scenes in which it can appear. For these reasons, the most general scenario of *markerless* pose estimation in monocular RGB images is often simplified by the introduction of constraints. The use of multiple cameras, indoor studios, artificial markers over body parts and depth sensors can all help to simplify the problem. Despite these strategies, in many cases images from single RGB cameras are the only ones available; in other cases, the imposition of constraints and intrusions can be difficult or undesirable, such as in outdoor environments, or in controlled areas such as medical facilities. Besides that, for applica-

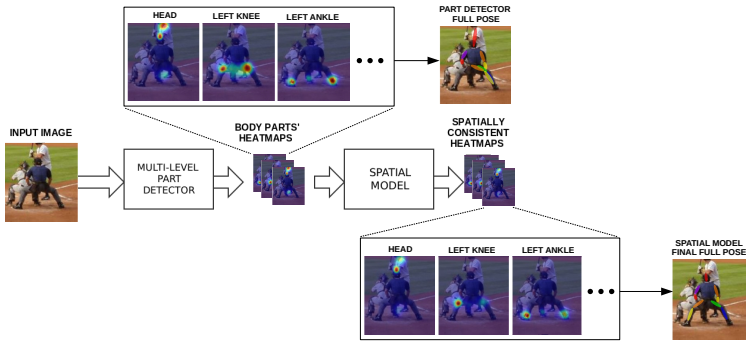


Figure 1: Our architecture, consisting of a part detector, which produces a heatmap for each body part, and a spatial model, which encourages spatial consistency over the detected body parts. A sample input image, along with sample heatmaps and full poses, illustrates how it can help with problems such as double-counting.

tions such as autonomous driving, the use of single RGB sensors greatly simplifies and reduces the cost of the data acquisition process, obviating the need for steps such as simultaneous calibration, image registration or synchronisation. Markerless pose estimation in monocular RGB images thus remains a key goal of current research.

Recently, important advances have been made by using deep learning to tackle the problem, greatly contributing to the improvement of human pose estimation in less-controlled scenarios (Elhayek et al., 2015). To our knowledge, the first effort in this direction was made by Taylor et al. (2010), who employed deep convolutional architectures for learning embeddings of images with people in similar poses, but with different clothes, background and other appearance changes. "Deep" human pose estimation methods have advanced the state of the art (see Sec. 2), but occlusions, cluttered scenes and unusual poses, which may lead to false positives and negatives when locating body parts, still represent challenges.

In this paper, we propose a deep CNN architecture, illustrated in Fig. 1, to tackle 2D markerless human pose estimation in monocular RGB images. Our architecture is composed of a multi-level body part detector and a fully-connected CRF model. Both modules are jointly trained end-to-end. We show that our detector, producing a holistic representation of the human body, benefits from the use of auxiliary parts (e.g. rigid parts) to locate joints. It is also noticeable that the CRF model, using the spatial relationships between body parts, learns which are the ones that exert more influence on each other, thus encouraging the generation of consistent poses. We evaluate our approach on both the MPII (Andriluka et al., 2014) and LSP (Johnson and Everingham, 2010) datasets, achieving competitive results and outperforming state-of-the-art methods in particular cases.

2. Related Work and Contribution

Recent years have seen outstanding improvements in 2D human pose estimation, driven by the introduction of CNN-based methods. Early works in this period have established two main lines of approach: the direct regression of joints' sparse coordinates (Toshev and Szegedy, 2014), and the dense regression of joints' locations over the image plane (Jain et al., 2014). The latter class of approaches has received greater attention from researchers due to some of its advantages, such as invariance to translation, multi-modality and a natural capability to handle smoother cost functions.

Many methods in this class have adopted exclusively bottom-up strategies, where joints are detected without any explicit use of prior knowledge or spatial reasoning about the human body

(Gkioxari et al., 2016; Rafi et al., 2016; Pfister et al., 2015; Belagiannis and Zisserman, 2017; Wei et al., 2016; Bulat and Tzimiropoulos, 2016). Other works employ different forms of ad-hoc, top-down refinements (Jain et al., 2014; Newell et al., 2016; Carreira et al., 2016; Hu and Ramanan, 2016; Lifshitz et al., 2016; Ke et al., 2018). Lastly, another group of approaches (Tompson et al., 2014; Chen and Yuille, 2014; Tompson et al., 2015; Pishchulin et al., 2016; Insafutdinov et al., 2016; Chu et al., 2016, 2017) perform structured predictions relying on part-based models (Fischler and Elschlager, 1973). Such models (Felzenszwalb and Huttenlocher, 2005; Yang and Ramanan, 2011), which are in fact instances of probabilistic graphical models such as MRFs and CRFs (Koller and Friedman, 2009; Lafferty et al., 2001), used to achieve state-of-the-art results on 2D pose estimation until the advent of CNN techniques. Our architecture is designed to jointly benefit from a bottom-up, CNN-based part detector and a part-based, fully-connected CRF model. We build upon knowledge and concepts from several lines of works in the literature, as we will now describe in more detail.

Multi-level appearance representations, which simultaneously gather coarse and fine visual cues, are well-known in the pose estimation literature (Pishchulin et al., 2013; Mikolajczyk et al., 2004). The underlying principle is that stronger and more diverse visual features from images can facilitate the location of joints. Concerning CNN-based approaches, local and global appearance models are employed in Fan et al. (2015), whilst in Belagiannis and Zisserman (2017), auxiliary body parts are used in the location of joints. In our work, we propose a simple weak annotation method to automatically derive the ground truth for the rigid parts and whole body from the manually annotated joints' coordinates.

Fully convolutional CNN architectures employing an efficient sliding-window technique (Sermanet et al., 2014; Long et al., 2015) appear in the context of pose estimation in Tompson et al. (2014), which has been followed by several other works. We use shared features as a **multitask learning** strategy to benefit from auxiliary body parts when locating the joints. Caruana (1998) mentioned that secondary tasks (e.g. detection of rigid parts and whole body) may produce an *inductive bias* towards the correct learning of a main task, and also act as regularisers in many cases. Li et al. (2015) apply a multitask approach in their CNN architecture for simultaneous regression and detection of body parts. By contrast, we handle multiple tasks in a cascade architecture, inspired by Dai et al. (2016)'s instance segmentation approach. **Cascaded CNN architectures**, such as Wei et al. (2016), have their foundations in cascaded or stacked classifiers (Heitz et al., 2009; Tu and Bai, 2010), in which classification outputs are sequentially refined through chains of classifiers.

Probabilistic graphical models for human pose estimation are explored jointly with CNNs in Tompson et al. (2014, 2015), where the authors introduced a method based on an MRF. In Chen and Yuille (2014), Pishchulin et al. (2016) and Insafutdinov et al. (2016), unaries and pairwise terms are explored in graphical model frameworks associated with deep architectures, but are not implemented in an end-to-end manner. Chu et al. (2017) achieved very good results by employing a CRF model on top of a pose estimation architecture (Newell et al., 2016). Our graphical model approach is closely related to Kiefel and Gehler (2014), in which the authors proposed a mean-field inference algorithm over a CRF model for human pose estimation using HOG features (Dalal and Triggs, 2005). Such algorithms were originally inspired by Krähenbühl and Koltun (2011), which showed how to perform efficient approximate inference over fully-connected CRF models for semantic segmentation. More recently, Zheng et al. (2015) showed how the mean-field algorithm could be formulated as a recurrent network for pixel-wise classification. However, it is important to note that the concept of a part does not exist in either Zheng et al. (2015) or Krähenbühl and Koltun (2011), preventing their direct application to part-based models.

In this paper, differently from previous approaches, we present a deep architecture to perform mean-field inference in part-based models applied to markerless human pose estimation. We make the following main contributions:

i) A multi-level Gaussian representation for the human body, and a novel, inexpensive and simple weakly-supervised methodology for generating corresponding ground-truth annotations.

ii) A novel framework for performing deep learning and inference in part-based models. To our knowledge, the presented method is the first to allow mean-field approximate inference over a loopy, fully-connected, part-based model using a deep end-to-end architecture. In our CRF, each body part corresponds to a distinct binary random field, and all parts together compose the part-based model (e.g. the human body). Pairwise functions are defined only between binary random variables in different random fields corresponding to *neighbouring* body parts: we dub these *inter-field* relations. Such relations are established according to the structure of the part-based model (e.g. fully-connected), differently from *intra-field* relations (i.e. between variables in the same field), as used in CRF models for image segmentation (Zheng et al., 2015; Krähenbühl and Koltun, 2011).

iii) Competitive results on two well-established benchmarks for 2D human pose estimation, the MPII (Andriluka et al., 2014) and the LSP (Johnson and Everingham, 2010) datasets, outperforming state-of-the-art methods in particular cases, e.g. unusual poses.

iv) Finally, although we present here the more general fully-connected setup, with joints' locations as features, our method allows for different part-based models' structures (e.g. tree-structured, star-structured, etc.), as well as for the use of multiple and different image features (e.g. colour, texture and depth). Thus, our framework may be applied to different deep part-based models for pose estimation, and also to other problems in which part-based models are employed, such as object co-detection (Hayder et al., 2014) and multi-person detection (Liu et al., 2016).

3. Our Approach

As shown in Fig. 1, we propose a deep neural network architecture that consists of two modules: a multi-level body part detector and a spatial model. The detector finds human body parts at multiple levels of granularity (joints, rigid parts and whole body), producing a dense heatmap for each part, whilst the spatial model encourages the prediction of consistent poses. We compose the two modules into a neural network and train it end-to-end using ground-truth heatmaps, constructed from manually-annotated positions in the case of the joints, and automatically generated with a simple, weakly supervised strategy for the rigid parts and the whole body (since they are not labelled on the employed datasets). The way in which we generate these heatmaps and the structure of the two network modules are described in the following subsections.

3.1. Multi-Level Gaussian Representation

We represent human pose using a set of oriented 2D Gaussian heatmaps, one for each joint (e.g. right shoulder) and rigid part (e.g. right forearm), and an additional one for the body as a whole. This approach builds upon recent successful approaches in the literature (Wei et al., 2016; Newell et al., 2016) that modelled joints alone as 2D Gaussians. The intuition behind adding rigid parts and the whole body to our representation is that in many images, they can be easier to see than the joints. Thus, these extra visual cues at multiple levels of granularity, i.e. semi-global and global (Pishchulin et al., 2013; Bourdev and Malik, 2009), act as inductive biases (Caruana, 1998) that allow us to better estimate the correct locations of the joints.

(a) (b) (c)

Figure 2: The 2D Gaussian of each body part is contained in a different heatmap, however for simplicity, all the Gaussians are superimposed together on the same image here. (a) 2D Gaussians over the joints. (b) 2D Gaussians over the rigid parts. (c) 2D Gaussians over the whole body.

Formally, our representation consists of $P = J + R + B$ body elements, namely J joints (numbered $1 \dots J$), R rigid parts (numbered $J+1 \dots J+R$) and one element ($B = 1$) for the whole body (numbered $J + R + B$). Each body element p is represented using a 2D Gaussian around its centre $\boldsymbol{\mu}_p = (i_p, j_p)$, with covariance matrix Σ_p, i.e. $G_p(i,j) = \frac{1}{\sqrt{2\pi|\Sigma_p|}} \exp\left(-\frac{1}{2}\begin{bmatrix} i-i_p \\ j-j_p \end{bmatrix}^\top \Sigma_p^{-1} \begin{bmatrix} i-i_p \\ j-j_p \end{bmatrix}\right)$. In this, (i, j) denotes a spatial position within the heatmap, with $i \in \{1, ..., H\}$ and $j \in \{1, ..., W\}$, where H and W are the heatmap height and width, respectively. The covariance matrix for a body element p can be decomposed as $\Sigma_p = R_p \begin{bmatrix} \sigma_{p,i}^2 & 0 \\ 0 & \sigma_{p,j}^2 \end{bmatrix} R_p^\top$, in which $\sigma_{p,i}$ and $\sigma_{p,j}$ respectively denote the standard deviations of the Gaussian along the principal axes of the body element, and R_p denotes the rotation from image space to body element space that defines the Gaussian's orientation. The multi-level body representation is illustrated in Fig. 2.

3.2. Ground-Truth Heatmap Generation

As mentioned above, to train our network we need ground-truth heatmaps for the body elements, but existing datasets have only been annotated with 2D joint locations (Johnson and Everingham, 2010; Andriluka et al., 2014; Gong et al., 2017). We thus propose a simple weakly supervised and part type-specific strategy of extending these to heatmaps:

Joints. Since joints have a limited spatial extent, we follow previous works (Wei et al., 2016; Newell et al., 2016) in modelling them as isotropic Gaussians (i.e. $\sigma_{p,i} = \sigma_{p,j}$ and $R_p = I$) that are centred at the ground-truth joint location and have a small standard deviation.

Rigid Parts. The centre $\boldsymbol{\mu}_p$ of a rigid part p is defined as the midpoint of the centres $\boldsymbol{\mu}_{p'}$ and $\boldsymbol{\mu}_{p''}$ of the joints it connects. We orient the Gaussian representing the part to align its i axis with the line connecting $\boldsymbol{\mu}_{p'}$ and $\boldsymbol{\mu}_{p''}$. We define $\sigma_{p,i}$ to be proportional to the Euclidean distance $||\boldsymbol{\mu}_{p'} - \boldsymbol{\mu}_{p''}||$, and set $\sigma_{p,j} = \kappa_p \sigma_{p,i}$, where κ_p is a part-specific anthropometric ratio (MSIS, 1995).

Body. The body centre is defined to be the mean of the annotated joint centres. Principal component analysis (PCA) of the joint centres is used to obtain the orientation of the body in the image plane. We define $\sigma_{p,i}$ and $\sigma_{p,j}$ to be proportional to the distance between the extreme projections of the joint centres onto the two principal axes.

Finally, we employ two sets of ground-truth heatmaps, one containing all the people in the scenes and the other containing only the person of interest. The second set is only used by the last module of the part detector architecture (Fig. 3). A sample ground-truth set is illustrated in Fig. 2.

3.3. Multi-Level Body Part Detector

This initial module of our network aims to simultaneously detect all body elements under consideration, and thereby refine the detection of the joints by making use of the additional visual cues

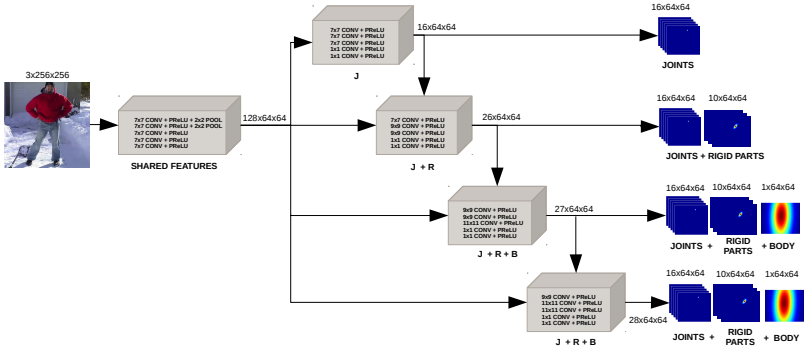


Figure 3: **Multi-Level Body Part Detector Basic Architecture**: a shared convolutional module produces CNN features that are used by subsequent modules to detect joints, rigid parts and the whole body. The initial detections of the joints (heatmaps) obtained by the module **J** – which can be seen as saliency maps (Itti et al., 1998) – are used to focus the *visual attention* (Borji and Itti, 2013) of the subsequent module **J+R**, which refines the previous predictions at the same time as it produces additional predictions for the rigid parts. The same principle is followed by the third module **J+R+B**, which refines the previous predictions and detects the whole body. The fourth module **J+R+B** is responsible for focusing on the person of interest, usually in the centre of the image. References to the backward pass of the network are not included for simplicity.

provided by the rigid parts and whole body. To achieve this, we adopt a cascade architecture where a common set of features are used to predict all body parts (see Fig. 3). This is reasonable due to the strong correlations between the visual appearance of the parts.

The proposed fully convolutional architecture produces one dense heatmap for each body part. Only two pooling layers are used over the network, reducing losses of spatial information. For our activation function, we use the Parametric ReLU (He et al., 2015). Another important point is the larger receptive field, which captures more contextual information around the joints.

The detector module as a whole implements a function of the form $f : \mathbb{R}^{H' \times W' \times 3} \times \mathbb{R}^n \mapsto \mathbb{R}^{H \times W \times P}$, which takes an RGB input image $\mathbf{Z}$ and a vector $\mathbf{w}$ of network weights, and yields an output tensor $\mathbf{X}$ that contains a dense heatmap for each of the P body elements under consideration. To learn $\mathbf{w}$, we train the network on a set of images $\{\mathbf{Z}^{(s)}\}$, with $s = 1, ..., S$. For each sample $\mathbf{Z}^{(s)}$ in the training set, we let $\mathbf{X}^{(s)} = f(\mathbf{Z}^{(s)}, \mathbf{w})$, and solve

$$\underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{S} \sum_{s=1}^{S} \sum_{p=1}^{P} \sum_{(i,j)} \mathcal{L}(\mathbf{X}_{i,j,p}^{(s)}; G_p(i,j)), \tag{1}$$

where $(i, j) \in H \times W$ denotes a spatial position within a heatmap; $\mathbf{X}_{i,j,p}^{(s)}$ denotes one value of the output tensor at a given coordinate; $G_p(i, j)$ is part p's 2D Gaussian, and $\mathcal{L}$ denotes mean squared error. Eq. 1 is minimized using stochastic gradient descent and, for each module in the cascade architecture, we use a similar loss function, including only the terms corresponding to the body elements detected by that module.

3.4. Fully-Connected Conditional Random Field

Our fully-connected CRF model is defined over the multi-level appearance representation as a framework for performing learning and inference on loopy part-based models for human pose estimation. By contrast with others in the literature, in our approach, not only all body parts within

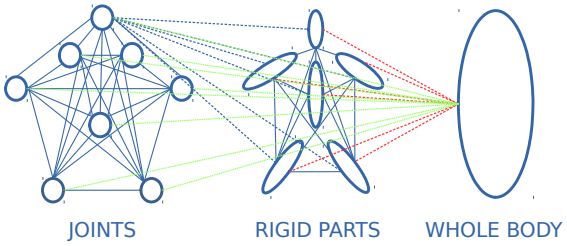


Figure 4: **Spatial model illustration**: The pairwise relations are illustrated as follows: **continuous blue lines** show *joint* $\leftrightarrow$ *joint* and *rigid part* $\leftrightarrow$ *rigid part* relations; **dashed blue lines** show *joint* $\leftrightarrow$ *rigid part* relations; **dashed red lines** show *rigid part* $\leftrightarrow$ *body* relations; **dotted green lines** show *joints* $\leftrightarrow$ *body* relations. Not all relations are illustrated for clarity. Best viewed in colour.

the same level of granularity are fully-connected by means of pairwise relations, but also all parts among all levels of granularity are connected, as illustrated in Fig. 4. Redundancy and the strong correlation between rigid parts and joints are positive characteristics in our model. The full connectivity allows each body part to receive messages from all others. This reinforces their consistent location, since ultimately, the posture of the whole body is taken into account in the message passing. Moreover, instead of imposing a fixed simpler structure *a priori* e.g. a tree, the strength of the pairwise relations and influence of the parts over each other, and consequently the underlying structure that emerges from that, is data-driven, i.e. learned from training data. Regarding the close relation between rigid parts and joints, as the former usually have a larger area in the images, they provide an inductive bias (Caruana, 1998) towards the correct location of the latter, since the joints are smaller and consequently more susceptible to being ambiguously or wrongly located, e.g. due to clutter or occlusion.

Formally, we introduce random binary variables $Y_{u,p}$ for $p = 1, ..., P$ and $u = 1, ..., H \times W$, each of which can take a value of either 0 or 1, respectively denoting the absence or presence of body part p at raster position u in an image $\mathbf{Z}$. Let $\mathbf{Y}$ be the vector composed of all such binary random variables $Y_{u,p}$. Let us now consider a factor graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with vertices $\mathcal{V} = \{\mathcal{V}_1, ..., \mathcal{V}_P\}$ divided in subsets $\mathcal{V}_p = \{Y_{u,p}\}$ corresponding to each one of the body parts; and edges $\mathcal{E} = (u, v)$, for $u \in \mathcal{V}_p$ and $v \in \mathcal{V}_{p'}$ with $p \neq p'$. In summary, an edge is established between every binary random variable of a given part and all the other random variables of all the other body parts in the appearance representation. Given the observable image $\mathbf{Z}$ and the vector $\mathbf{Y}$, the pair $(\mathbf{Y}, \mathbf{Z})$ can be modelled as a CRF characterised by a Gibbs distribution $P(\mathbf{Y} = \mathbf{y}|\mathbf{Z})$. For convenience of notation, we will suppress the conditioning in the following equations. Thus, the energy of a given configuration $\mathbf{y}$ is given by

$$E(\mathbf{y}) = \sum_{p=1}^{P} \left(\sum_{u \in \mathcal{V}_p} \phi(y_{u,p}) + \sum_{p'=1}^{P} \sum_{u \in \mathcal{V}_p} \sum_{v \in \mathcal{V}_{p'}} \psi(y_{u,p}, y_{v,p'}) \right), \tag{2}$$

where the unary potentials $\phi(y_{u,p})$ give the likelihood for the presence of a part p in a given location u, and the pairwise potentials $\psi(y_{u,p}, y_{v,p'})$ measure the likelihood for the simultaneous location of parts p and p' in the corresponding positions u and v, respectively. In the proposed model, the unary terms correspond to the energies provided by the CNN-based body part detector (Sec. 3.3).

The pairwise potentials are defined by weighted Gaussian kernels as

$$\psi(y_{u,p}, y_{v,p'}) = \sum_{k=1}^{K} \mathbf{w}_{p,p'}^{(k)} \boldsymbol{\mu}_{p,p'}^{(k)}(y_{u,p}, y_{v,p'}) \mathbf{k}_{p,p'}^{(k)}(\mathbf{f}_{u,p}, \mathbf{f}_{v,p'}),^{1} \tag{3}$$

where we have a linear combination of Gaussian kernels $\mathbf{k}_{p,p'}^{(k)}(.)$, weighted by $\mathbf{w}_{p,p'}^{(k)}$, and the compatibility functions $\boldsymbol{\mu}_{p,p'}^{(k)}(.)$. Each Gaussian kernel is applied on feature vectors $\mathbf{f}_{u,p}$ and $\mathbf{f}_{v,p'}$ from the image, which may encode information such as the location, colour or intensity of the parts. The weights measure the relative importance of each kernel. Finally, the compatibility function is a 2×2 matrix, as it weights how one binary label assignment for one part influences the assignment for the other part. It is important to notice that there are a different Gaussian kernel and compatibility functions between each pair of parts p and p', because of the different ways that different parts influence each other. This differs from the DenseCRF approach for semantic segmentation (Krähenbühl and Koltun, 2011; Zheng et al., 2015), where a single compatibility function and Gaussian kernel are used for the entire model. In our case, we explicitly model the dependencies between different parts in the part-based model, e.g. human body parts. Although our method allows for multiple kernels between each pair of parts p and p', in the current formulation we only use the location of the parts as a feature ($K = 1$), thus the feature vectors $\mathbf{f}_{u,p}, \mathbf{f}_{v,p'}$ correspond to the 2D positions of parts p and p', respectively. The Gaussian kernel is then defined as

$$\mathbf{k}_{p,p'}^{(k)}(\mathbf{f}_{u,p}, \mathbf{f}_{v,p'}) = \exp\Big\{ -\frac{1}{2}\big[(\mathbf{f}_{u,p} - \mathbf{f}_{v,p'}) - \bar{\mathbf{x}}_{p,p'}\big]^T \Sigma_{p,p'}^{-1} \big[(\mathbf{f}_{u,p} - \mathbf{f}_{v,p'}) - \bar{\mathbf{x}}_{p,p'}\big]\Big\}, \tag{4}$$

where the vector $\bar{\mathbf{x}}_{p,p'}$ and the matrix $\Sigma_{p,p'}$ are, respectively, the 2D mean displacement and the covariance between the locations of parts p and p'. Thus the Gaussian kernel has maximum value when the distance between two parts is equal to their mean displacement and exponentially decays when their offset moves away from it.

3.5. Mean-Field Inference as a RNN in a Part-Based Model

Our final predicted human pose is the Maximum a Posteriori (MAP) estimate of the CRF defined in Eq. 2. This corresponds to the assignment $\mathbf{y}^*$ that minimises the energy $E(\mathbf{y})$. Exact inference over this loopy fully-connected CRF is intractable. As a result, we perform approximate mean-field inference in which a simpler distribution, expressed as a product of independent marginals, $Q(\mathbf{Y}) = \prod_u Q(Y_{u,p})$, is used to approximate the real distribution, $P(\mathbf{Y})$. The KL-divergence between $P(\mathbf{Y})$ and $Q(\mathbf{Y})$ is then iteratively minimised. The mean-field update equation (Koller and Friedman, 2009) for our fully-connected CRF model is given by

$$Q(y_{u,p}) \propto \exp\Big\{ -\phi(y_{u,p}) - \sum_{p'=1}^{P} \sum_{k=1}^{K} \mathbf{w}_{p,p'}^{(k)} \sum_{l' \in \{0,1\}} \boldsymbol{\mu}_{p,p'}^{(k)}(y_{u,p}, l') \sum_{v \in \mathcal{V}_{p'}} \mathbf{k}_{p,p'}^{(k)}(\mathbf{f}_{u,p}, \mathbf{f}_{v,p'}) Q(l') \Big\}. \tag{5}$$

Eq. 5 contains a highly costly message-passing operation performed in the pairwise terms calculation. The definition of such pairwise relations as Gaussian kernels allows their computation by means of a convolution operation, which can be efficiently performed using a permutohedral lattice (Adams et al., 2010), reducing the complexity of the message-passing step from quadratic to linear with respect to the number of random variables.

It has recently been shown that mean-field iterative updates are differentiable (Zheng et al., 2015; Arnab et al., 2016). Consequently, the mean-field inference algorithm can be represented

1. The main functions and variables are written in bold here to facilitate reading.

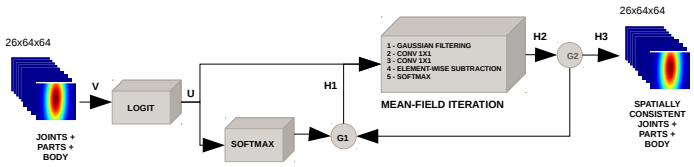


Figure 5: **Spatial Model Architecture**: the **mean-field iteration** module executes the mean-field update for T iterations. The additional modules pre-process the input unaries. Gates G1 and G2 switch outputs according to Eq. 6. References to the backward pass of the network are not included for simplicity.

as an RNN, where the estimate of $Q(\mathbf{Y})$ is refined at each time step. This allows us to learn the parameters of our spatial model and the underlying CNN jointly via end-to-end training. As shown in Fig. 5, our RNN receives as inputs the unary potentials from the CNN-based part detector, and produces spatially consistent heatmaps (marginal distributions) as outputs. Its loss function is the same given by Eq. 1.

To define the overall behavior of the network, let us consider the unary potential $U = -\phi(y_{u,p}) = \text{logit}(V)$. We apply the logit(.) function (inverse Sigmoid) over the body part detector outputs to compute the log-likelihood scores corresponding to the unary energies. Once the unaries are obtained, an initial normalisation step is executed by the softmax(.) function. The Softmax and Logit modules have their error differentials trivially back-propagated. Additionally, let us define an estimation of the marginal probabilities Q from the previous step. And finally, the forward-pass of the mean-field iteration module can be denoted by the function $f_\theta(U, Q)$, where the vector $\theta = \{\mathbf{w}^{(k)}_{p,p'}, \boldsymbol{\mu}^{(g)}_{p,p'}(l, l')\}$ represents the *learnable* CRF parameters. In this context, the following equations define the RNN behavior, where T is the total number of iterations:

$$
\begin{aligned}
H_1(t) &= \begin{cases} \text{softmax}(U), & t = 0 \\ H_2(t-1), & 0 < t \leq T, \end{cases} \\
H_2(t) &= \begin{cases} f_\theta(U, H_1(t)), & 0 \leq t \leq T, \end{cases} \\
H_3(t) &= \begin{cases} 0, & 0 \leq t < T \\ H_2(t), & t = T. \end{cases}
\end{aligned}
\tag{6}
$$

Internally, the sequence of operations executed by the mean-field iteration module (i.e. *Gaussian filtering*, *convolutions*, *element-wise subtraction* and *Softmax*) correspond to the the mean-field steps for CRF models defined with a unary term, and pairwise terms based on Gaussian kernels (Krähenbühl and Koltun, 2011; Kiefel and Gehler, 2014; Zheng et al., 2015).

4. Experiments and Discussion

We evaluate the performance of the part detector and the spatial model, separately and together. Following Bulat and Tzimiropoulos (2016); Wei et al. (2016); Insafutdinov et al. (2016); Pishchulin et al. (2016) and Belagiannis and Zisserman (2017), we have trained our network by adding the MPII training set to the LSP and LSP extended training sets. Tests were done on the LSP and MPII test sets, using PCK (Yang and Ramanan, 2011) and PCKh (Andriluka et al., 2014) metrics, respectively. In all experiments, we adopted the following parameters: mini-batch equal to 10; Adam optimizer (Kingma and Ba, 2015) with learning rate equal to 10^{-5}; weight decay of 5×10^{-4} and no other form of regularization; and network weights initialized using the robust initialization proposed by He et al. (2015). We cropped a 256×256 image area, keeping the person of interest in

the central position, reinforcing the focus on it using an extra input channel containing a central and fixed size 2D Gaussian (standard deviation equal to 15 pixels), which is concatenated to the input channels of the final module of the network. Regarding data augmentation, we randomly apply the following over the images: scaling $[0.5, 1.5]$, rotation $[-45°, 45°]$, horizontal flipping and RGB jittering (Krizhevsky et al., 2012). We evaluate variants of our architecture and the contribution of the CRF model over a single scale in a validation set. At test time, we perform predictions over scales in $[0.5, 1.5]$, equally spaced with a step of 0.1, and sum them to obtain the final estimations. The architecture was implemented using the Caffe framework (Jia et al., 2014) and the experiments ran on an NVIDIA Titan X.

Multi-Level Body Part Detector. To evaluate variants of the basic architecture illustrated in Fig. 3, and support our design choices, we have trained different arrangements of modules, with the parameters afore detailed. They are defined by bold capital letters, e.g. **J** represents one module that locates joints, **(J)(J+R)** represents two modules which predict joints and joints along with rigid parts, respectively, and **4×(J+R+B)** denotes four modules predicting joints, rigid parts and whole body. It is important to highlight that, except when stated, the variants are cascaded, and every module has a loss weight equal to 1. A summary of the variants is listed in Tab. 1, with their scores and reference names. In what follows, we analyse some of the evidence observed. Initially, the sets of architectures **A** and **B** not only show the benefits of multiple modules ($\mathbf{B_1}$ vs. $\mathbf{A_1}$), but also show that cascaded modules with intermediate losses ($\mathbf{B_3}$) are preferable to either sequential or cascade networks of the same capacity, but with just one final loss function ($\mathbf{B_1}$ and $\mathbf{B_2}$). The set **C** corresponds to our basic architecture, and shows evidence that the auxiliary body parts somehow improve the performance ($\mathbf{C_2}$ vs. $\mathbf{B_3}$), conditioned to the use of weights ($\mathbf{C_2}$ vs. $\mathbf{C_1}$) attributing more importance to the joints and to final modules in the overall loss. To evaluate the effect of having the extra parts gradually added through the architecture, we employed sets **D** and **E**. The results suggest that having all modules predicting the same multi-level representation is preferable ($\mathbf{D_1}$ vs. $\mathbf{C_2}$). Besides this, the prediction of the whole body does not seem to help in the location of the joints ($\mathbf{E_1}$). We hypothesize that this happens because the person of interest is already centralized in most images of the LSP dataset. Finally, we have evaluated the addition of extra modules in sets **F** and **G**, finding the overall best performing architecture as $\mathbf{F_1}$. We also tried other experiments over the sets, such as inverting the order of the modules in $\mathbf{C_1}$ and $\mathbf{C_2}$, adding an extra background heatmap in the final modules and changing combinations of weights, but none of these improved the final performance.

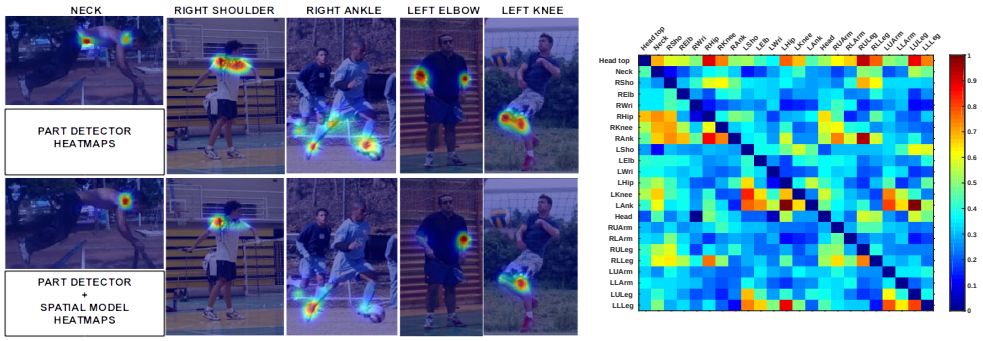
Reference name	Architecture	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
$\mathbf{A_1}$	**J**	91.2	78.7	64.7	54.3	71.8	61.7	50.0	67.5
$\mathbf{B_1}_{\text{SEQUENTIAL / WITH\_JUST\_FINAL\_LOSS}}$	**4×(J)**	92.8	82.8	72.0	63.6	80.7	74.0	63.8	75.7
$\mathbf{B_2}_{\text{WITH\_JUST\_FINAL\_LOSS}}$	**4×(J)**	93.1	82.1	70.2	63.0	79.2	72.4	60.0	74.3
$\mathbf{B_3}$	**4×(J)**	93.5	82.6	72.9	66.3	81.3	75.0	65.6	76.8
$\mathbf{C_1}$	**(J)+(J+P)+(J+P+B)+(J+P+B)**	93.4	81.7	71.2	63.2	80.4	75.2	69.0	76.3
$\mathbf{C_2}_{\text{J\_LOSSES\_WEIGHTED } \{1,2,3,12\}^\dagger}$	**(J)+(J+P)+(J+P+B)+(J+P+B)**	94.0	83.2	**75.0**	66.0	81.2	76.4	71.0	78.1
$\mathbf{D_1}_{\text{J\_LOSSES\_WEIGHTED } \{1,2,3,12\}^\dagger}$	**4×(J+P)**	94.2	**84.2**	74.5	67.2	81.7	77.0	71.9	78.7
$\mathbf{E_1}_{\text{J\_LOSSES\_WEIGHTED } \{1,2,3,12\}^\dagger}$	**4×(J+P+B)**	92.8	79.8	69.9	61.3	78.7	71.8	67.3	74.5
$\mathbf{F_1}_{\text{J\_LOSSES\_WEIGHTED } \{1,2,3,12,24\}^\dagger}$	**5×(J+P)**	**94.3**	83.9	73.3	**68.0**	**83.4**	**78.7**	**72.8**	**79.2**
$\mathbf{G_1}_{\text{J\_LOSSES\_WEIGHTED } \{1,2,3,12,24,48\}^\dagger}$	**6×(J+P)**	92.8	82.4	73.6	67.2	81.1	76.1	70.4	77.6

† Sets of relative weights of the modules losses.

Table 1: Comparison of PCK@0.2 on the LSP dataset.

Reference name	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
$\mathbf{F_1}$	94.3	83.9	73.3	68.0	**83.4**	**78.7**	72.8	79.2
$\mathbf{F_1}$+SM	**94.3**	**84.2**	**74.5**	**69.1**	83.2	78.5	**73.2**	**79.6**

Table 2: Comparison of PCK@0.2 on the LSP dataset. SM = Spatial Model.



(a) (b)

Figure 6: (a) **Heatmaps from the LSP dataset.** The first row shows heatmaps obtained from the part detector ($\mathbf{F_1}$), while the second row shows heatmaps from the part detector plus the spatial model ($\mathbf{F_1}$+**SM**). (b) **Learned Mean-field parameters:** the higher the compatibility value, the higher the influence between the parts. We can notice small clusters of high influence, particularly between parts on the same side of the body (left/right), which explains how we migitate double-counting.

Conditional Random Field. Here we evaluated the contribution of the fully-connected CRF. The compatibility matrices were initialized as $\boldsymbol{\mu}^{(k)}_{p,p'} = \left[\begin{smallmatrix} 0 & 0 \\ 0 & 0 \end{smallmatrix}\right]$, denoting that the influence between the different body parts is initially null and will be learned during training. The kernel weights are set to $\mathbf{w}^{(k)}_{p,p'} = 1$, once we employ a single Gaussian kernel with spatial features (Eq. 4) in the present basic formulation of the model. We adopted the best performing architecture from Tab. 1 ($\mathbf{F_1}$) and compared its results with the ones obtained when the spatial model was employed in the second half of the training ($\mathbf{F_1}$+**SM**). Tab. 2 shows the improvements obtained over the part detector performance. The current formulation using only spatial information was already beneficial, as shown by the samples of body part heatmaps obtained with and without the use of the spatial model, in Fig. 6a. We can observe that in fact the heatmaps are refined when prior knowledge about the body is taken into consideration. The learned mean-field parameters are shown in Fig. 6b, i.e. the compatibility values, since the kernel weights were kept fixed due to the single spatial kernel employed. The learned compatibility values $\boldsymbol{\mu}^{(k)}_{p,p'}$ of all the pairwise matrices represent how much each body part on the horizontal axis influences all other parts on the vertical axis.

LSP dataset. Tab. 3 shows our final results on the LSP test set, after 200 epochs, along with comparable approaches from the literature. We have competitive results in comparison to the state-of-the-art. Moreover, when the presence of false positives generated by clutter or by unusual poses prejudice the bottom-up methods, we may outperform such approaches, as illustrated in Fig. 7. The stronger inductive bias towards the correct positions of the joints, obtained with the use of the rigid parts in our multi-level representation, along with the high level of redundancy of the fully-connected CRF, facilitates the correct estimations in these cases.

MPII dataset. Tab. 4 shows results on the MPII test set, in this case employing just the MPII training data for 250 epochs. We have again shown competitive scores, and an ability to perform well in the presence of occlusions and unusual poses (see Fig. 8a). See failure cases in Fig. 8b.

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
Chu et al.	98.1	93.7	89.3	86.9	93.4	94.0	92.5	92.6
Wei et al.	97.8	92.5	87.0	83.9	91.5	90.8	89.9	90.5
Our model	97.6	92.6	86.5	83.2	91.7	91.0	89.1	90.2
Insafutdinov et al.	97.4	92.7	87.5	84.4	91.5	89.9	87.2	90.1
Pishchulin et al.	97.0	91.0	83.8	78.1	91.0	86.7	82.0	87.1
Belagiannis and Zisserman	95.2	89.0	81.5	77.0	83.7	87.0	82.8	85.2

Table 3: Comparison of PCK@0.2 score on the LSP test set. Only methods trained on MPII, LSP and LSP extended datasets jointly.

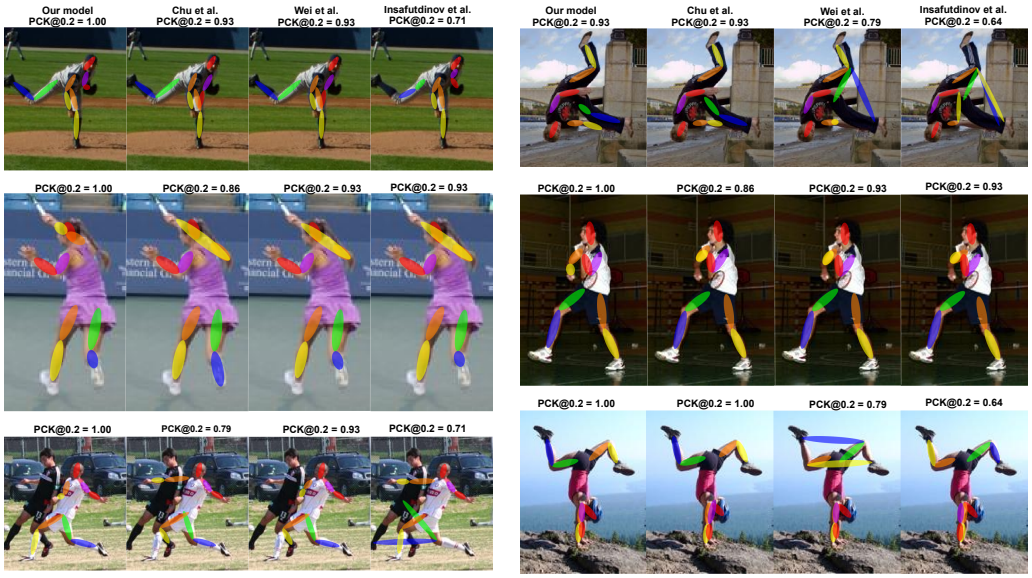


Figure 7: Cases from LSP in which we outperformed state-of-the-art bottom-up methods prejudiced by false positives caused by clutter and unusual poses.

Overall Analysis. Our method is able to cope with challenging poses, outperforming state-of-the-art methods in particular cases. From the analysis of our fully-connected CRF model, summarized in Fig. 6, we can observe that the spatial model contributes mostly to the distinction between left and right parts of the body, whatever ambiguity is generated by clutter, occlusions or unusual poses i.e. upside-down postures, as illustrated in Fig. 7. Such uncertainties are mitigated by data-driven compatibility matrix values (Fig. 6b), which show that parts on the same side of the body exert strong influence over each other. High compatibility values between parts that would not be directly connected in simpler tree-structured models, e.g. hips and ankles, also show that the fully-connected model is learning more complex relations between the body parts. Although we have presented a basic and general formulation of our spatial model, based only on the 2D locations of the parts, it has already shown marginal improvement over the part detector alone (Tab. 2). Additional features, such as the colours of mirrored body parts, may be naturally incorporated, potentially improving even more the performance in relation to the bottom-up part detector. Finally, in our experiments, the CRF model achieved convergence in a small number of iterations ($T = 3$) for the recurrent mean-field update, as similarly reported by Zheng et al. (2015). Despite that, training and testing with the fully-connected model are costly. With our best-performing model (Tabs. 1 and 2), when the CRF model is added to the CNN part-detector, training time per image increases 6

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
Ke et al.	98.5	96.8	92.7	88.4	90.6	89.3	86.3	92.1
Chu et al.	98.5	96.3	91.9	88.1	90.6	88.0	85.0	91.5
Newell et al.	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9
Bulat and Tzimiropoulos	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7
Insafutdinov et al.	96.8	95.2	89.3	84.4	88.4	83.4	78.0	88.5
Our model	97.7	95.0	88.1	83.4	87.9	82.1	78.7	88.1
Rafi et al.	97.2	93.9	86.4	81.3	86.8	80.6	73.4	86.3
Gkioxari et al.	96.2	93.1	86.7	82.1	85.2	81.4	74.1	86.1
Lifshitz et al.	97.8	93.3	85.7	80.4	85.3	76.6	70.2	85.0
Hu and Ramanan	95.0	91.6	83.0	76.6	81.9	74.5	69.5	82.4
Tompson et al.	96.1	91.9	83.9	77.8	80.9	72.3	64.8	82.0
Carreira et al.	95.7	91.7	81.7	72.4	82.8	73.2	66.4	81.3
Tompson et al.	95.8	90.3	80.5	74.3	77.6	69.7	62.8	79.6
Pishchulin et al.	74.3	49.0	40.8	34.1	36.5	34.4	35.2	44.1

Table 4: Comparison of PCKh@0.5 score on the MPII test set. Only methods trained on MPII dataset.

(a) (b)

Figure 8: (a) Sample images from the MPII dataset showing contexts in which we successful handle challenging cases, such as occlusions. (b) Sample images from the MPII dataset in which the method could not overcome the ambiguities in the poses.

times, from 0.3s to 1.8s, approximately, while testing time per image increases 5 times, from 1.5s to 7.5s, approximately (testing is performed over multiple scales, as previously mentioned in this section). For the final models, the total training time was approximately 7 days. As an alternative, this high cost might be diminished by the elimination of *weak* pairwise relations, i.e. the ones with small values in the compatibility matrix (Fig. 6b). In this case, the fully-connected model would function as an intermediate step towards a simpler yet still data-driven model structure.

5. Conclusions

In this paper, we have introduced a novel framework composed of a multi-level appearance representation of the human body, along with a loopy fully-connected part-based model, conveying information about the spatial structure of the body. We proposed a cascade CNN body part detector to obtain our appearance representation, and an RNN to perform mean-field inference on the part-based model, defined as a binary CRF. We evaluate the components of our architecture, showing that the multi-level representation, composed of body parts with different granularities, as well as the spatial model, facilitate the location of the body joints. Our experiments on the MPII and LSP benchmarks have shown competitive results. In particular cases, when the presence of false positives greatly prejudices exclusive bottom-up strategies, we outperformed the state-of-the-art methods. The current deep learning framework may be naturally extended to other problems, such as object detection and per-pixel labelling of objects parts, and also by the addition of extra kernels, based on other features such as colour. As a further improvement, we intend to generalize our spatial model to handle multimodal distributions, a limitation imposed on the current formulation by the use of Gaussian pairwise kernels, which are efficient to compute (Adams et al., 2010).

Acknowledgments

Rodrigo Andrade de Bem is a CAPES Foundation scholarship holder (Process no: 99999.013296/2013-02, Ministry of Education, Brazil). This work was done in the University of Oxford and supported by the EPSRC, ERC grant ERC-2012-AdG 321162-HELIOS, EPSRC grant Seebibyte EP/M013774/1 and EPSRC/MURI grant EP/N019474/1.

References

Andrew Adams, Jongmin Baek, and Myers Abraham Davis. Fast high-dimensional filtering using the permutohedral lattice. In *CGF*, 2010.

Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014.

Anurag Arnab, Sadeep Jayasumana, Shuai Zheng, and Philip Torr. Higher order conditional random fields in deep neural networks. In *ECCV*, 2016.

Vasileios Belagiannis and Andrew Zisserman. Recurrent human pose estimation. In *FG*, 2017.

Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *TPAMI*, 2013.

Lubomir Bourdev and Jitendra Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009.

Adrian Bulat and Georgios Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *ECCV*, 2016.

Joao Carreira, Pulkit Agrawal, Katerina Fragkiadaki, and Jitendra Malik. Human pose estimation with iterative error feedback. In *CVPR*, 2016.

Rich Caruana. Multitask learning. In *Learning to learn*, pages 95–133. Springer, 1998.

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

Xiao Chu, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Crf-cnn: Modeling structured information in human pose estimation. In *NIPS*, 2016.

Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. In *CVPR*, 2017.

Jifeng Dai, Kaiming He, and Jian Sun. Instance-aware semantic segmentation via multi-task network cascades. In *CVPR*, 2016.

Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

A Elhayek, E. de Aguiar, A. Jain, J. Tompson, L. Pishchulin, M. Andriluka, C. Bregler, B. Schiele, and C. Theobalt. Efficient ConvNet-based Marker-less Motion Capture in General Scenes with a Low Number of Cameras. In *CVPR*, 2015.

Xiaochuan Fan, Kang Zheng, Yuewei Lin, and Song Wang. Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. In *CVPR*, 2015.

Pedro F Felzenszwalb and Daniel P Huttenlocher. Pictorial structures for object recognition. *IJCV*, 2005.

Martin A Fischler and Robert A Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computers*, 1973.

Georgia Gkioxari, Alexander Toshev, and Navdeep Jaitly. Chained predictions using convolutional neural networks. In *ECCV*, 2016.

Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *CVPR*, 2017.

Zeeshan Hayder, Mathieu Salzmann, and Xuming He. Object co-detection via efficient inference in a fully-connected crf. In *ECCV*, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015.

Geremy Heitz, Stephen Gould, Ashutosh Saxena, and Daphne Koller. Cascaded classification models: Combining models for holistic scene understanding. In *NIPS*, 2009.

Peiyun Hu and Deva Ramanan. Bottom-up and top-down reasoning with hierarchical rectified gaussians. In *CVPR*, 2016.

Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *ECCV*, 2016.

Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *TPAMI*, 1998.

Arjun Jain, Jonathan Tompson, Mykhaylo Andriluka, Graham W. Taylor, and Christoph Bregler. Learning Human Pose Estimation Features with Convolutional Networks. In *ICLR*, 2014.

Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv:1408.5093*, 2014.

Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010.

Lipeng Ke, Ming-Ching Chang, Honggang Qi, and Siwei Lyu. Multi-scale structure-aware network for human pose estimation. *arXiv preprint arXiv:1803.09894*, 2018.

Martin Kiefel and Peter Vincent Gehler. Human pose estimation with fields of parts. In *ECCV*, 2014.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, 2001.

Sijin Li, Zhi-Qiang Liu, and Antoni B. Chan. Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network. *IJCV*, 2015.

Ita Lifshitz, Ethan Fetaya, and Shimon Ullman. Human pose estimation using deep consensus voting. In *ECCV*, 2016.

Jingjing Liu, Quanfu Fan, Sharath Pankanti, and Dimitris N Metaxas. People detection in crowded scenes by context-driven label propagation. In *WACV*, 2016.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

Krystian Mikolajczyk, Cordelia Schmid, and Andrew Zisserman. Human detection based on a probabilistic assembly of robust part detectors. *ECCV*, 2004.

Thomas B Moeslund, Adrian Hilton, Volker Krüger, and Leonid Sigal. *Visual analysis of humans*. Springer, 2011.

MSIS. Space flight human-system standard volume 1. Technical Report NASA-STD-3001, National Aeronautics and Space Administration - NASA, 1995.

Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked Hourglass Networks for Human Pose Estimation. In *ECCV*, 2016.

Tomas Pfister, James Charles, and Andrew Zisserman. Flowing convnets for human pose estimation in videos. In *ICCV*, 2015.

Leonid Pishchulin, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. Strong appearance and expressive spatial models for human pose estimation. In *ICCV*, 2013.

Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *CVPR*, 2016.

Umer Rafi, Ilya Kostrikov, Juergen Gall, and Bastian Leibe. An efficient convolutional network for human pose estimation. In *BMVC*, 2016.

Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *ICLR*, 2014.

Graham W Taylor, Rob Fergus, George Williams, Ian Spiro, and Christoph Bregler. Pose-sensitive embedding by nonlinear nca regression. In *NIPS*, 2010.

Jonathan Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation. In *NIPS*, 2014.

Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christopher Bregler. Efficient Object Localization Using Convolutional Networks. In *CVPR*, 2015.

Alexander Toshev and Christian Szegedy. DeepPose: Human Pose Estimation via Deep Neural Networks. In *CVPR*, 2014.

Zhuowen Tu and Xiang Bai. Auto-context and its application to high-level vision tasks and 3d brain image segmentation. *PAMI*, 2010.

Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016.

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

Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip Torr. Conditional random fields as recurrent neural networks. In *ICCV*, 2015.