Multi-Context Attention for Human Pose Estimation

Xiao Chu[1] * Wei Yang[1] * Wanli Ouyang[1,4] Cheng Ma[2] Alan L. Yuille[3] Xiaogang Wang[1]
[1] The Chinese University of Hong Kong, Hong Kong SAR, China
[2] Tsinghua University, Beijing, China
[3] Johns Hopkins University, Baltimore, USA
[4] The University of Sydney, Sydney, Australia
[1]{xchu, wyang, wlouyang, xgwang}@ee.cuhk.edu.hk
[2]macheng13@mails.tsinghua.edu.cn [3]alan.yuille@jhu.edu

Abstract

In this paper, we propose to incorporate convolutional neural networks with a multi-context attention mechanism into an end-to-end framework for human pose estimation. We adopt stacked hourglass networks to generate attention maps from features at multiple resolutions with various semantics. The Conditional Random Field (CRF) is utilized to model the correlations among neighboring regions in the attention map. We further combine the holistic attention model, which focuses on the global consistency of the full human body, and the body part attention model, which focuses on detailed descriptions for different body parts. Hence our model has the ability to focus on different granularity from local salient regions to global semantic-consistent spaces. Additionally, we design novel Hourglass Residual Units (HRUs) to increase the receptive field of the network. These units are extensions of residual units with a side branch incorporating filters with larger receptive field, hence features with various scales are learned and combined within the HRUs. The effectiveness of the proposed multi-context attention mechanism and the hourglass residual units is evaluated on two widely used human pose estimation benchmarks. Our approach outperforms all existing methods on both benchmarks over all the body parts. Code has been made publicly available.

1. Introduction

Human pose estimation is a challenging task in computer vision due to the articulation of body limbs, self occlusion, various clothing, and foreshortening. Significant improvements have been achieved by Convolutional Neural Networks (ConvNets) [37, 38, 9, 39, 36, 28]. However, for cluttered background with objects which are similar to body parts or limbs, or body parts with heavy occlusion,

*The first two authors contribute equally to this work.

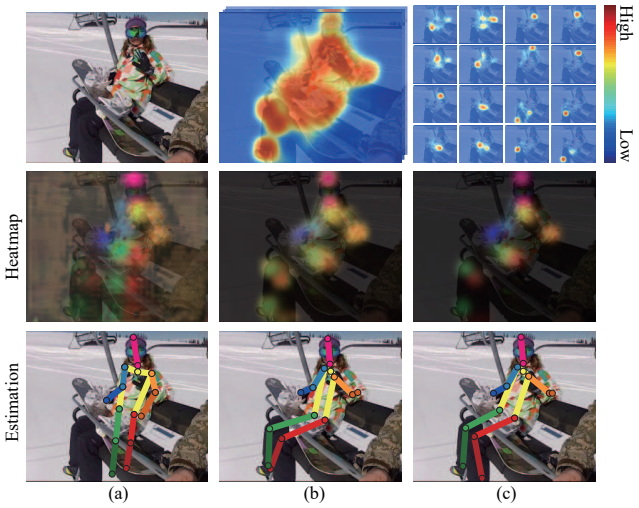

Figure 1. **Motivation**. The 1st row shows the input image, the holistic attention maps, and the part attention maps. The 2nd row shows the predicted heatmaps for part locations, where different colors correspond to different body parts. The 3rd row visualizes the predicted poses. We observe that (a) ConvNets may produce erroneous estimations due to cluttered background and self-occlusion. (b) Visual attention provides an explicit way to model spatial relationships among human body parts, which is more robust. (c) Part attention maps can help further refine the part locations by addressing the double counting problem.

ConvNets may have difficulty to locate each body part correctly, as demonstrated in Fig. 1 (a). In the literature, the combination of multiple contextual information has been proved essential for vision tasks such as image classification [25], object detection [15, 14, 49] and human pose estimation [33, 36]. Intuitively, larger context region captures global spatial configurations of object, while smaller context region focuses on local part appearance. However, previous works usually use manually designed multi-context representations, *e.g.*, multiple bounding boxes [33] or multiple image crops [25], and hence lack of flexibility and di-

versity for modeling the multi-context representations.

Visual attention is an essential mechanism of the human brain for understanding scenes effectively. In this work, we propose to generate contextual representations with an attention scheme. Instead of defining regions of interest manually by a set of rectangle bounding boxes, the attention maps are generated by an attention model, which depend on image features, and provide a principled way to focus on target regions with variable shapes. For example, an attention map focusing on the human body is shown in Fig. 1(b). It helps recover the missing body parts (*e.g.*, *legs*), and distinguishes the ambiguous background. This allows the diversity of context to be increased, and so contextual region could be better adapted to each image. Furthermore, instead of adopting the spatial Softmax normalization widely used in conventional attention schemes[47, 41, 46, 26], we design a novel attention model based on Conditional Random Fields, which is better in modeling the spatial correlations among neighboring regions.

The combination of multiple contextual information has been proved effective for various vision tasks [48, 15, 33, 13, 34]. To use the attention mechanism to guide multi-contextual representation learning, we adopt the stacked hourglass network structure [28], which provides an ideal architecture to build a multi-context attention model. In each hourglass stack, features are pooled down to a very low resolution, then are upsampled and combined with high-resolution features. This structure is repeated for several times to gradually capture more global representations. Within each hourglass stack, we first generate *multi-resolution* attention maps from features of different resolutions. Secondly, we generate attention maps for multiple hourglass stacks, which results in *multi-semantics* attention maps with various levels of semantic meaning. Since these attention maps capture the configuration of the full human body, they are referred to as holistic attention models.

While the holistic attention model is robust to occlusions and cluttered background, it lacks of precise description for different body parts. To overcome this limitation, we design a hierarchical visual attention scheme, which zooms in from holistic attention model to each body part, namely the *part attention model*. This is helpful for precise localization of the body parts, as shown in Fig. 1 (c).

Additionally, we introduce a novel "Hourglass Residual Units" as a replacement for the residual unit [19] in our network. It incorporates the expressive power of multi-scale features while preserving the benefit of residual learning. It also enables deep networks to have a faster growth of receptive field, which is essential for accurately locating body parts. When using these units within the "macro" hourglass network, we obtain a *nested* hourglass architecture.

We show the effectiveness of the proposed end-to-end differentiable framework on two broadly used human pose estimation benchmarks, *i.e.*, MPII Human Pose dataset [1] and the Leeds Sports Dataset [23]. Our approach outperforms all the previous methods on both benchmarks for all the body parts. Code has been made publicly available at https://github.com/bearpaw/pose-attention. The main contributions of this work are three folds:

- We propose to use visual attention mechanism to automatically learn and infer the contextual representations, driving the model to focus on region of interest. We tailor the attention scheme for human pose estimation by introducing CRFs to model the spatial correlations among neighborhood joints.
- We use multi-context attention to make the model more robust and more accurate.
- We propose a generic hourglass residual unit (HRU), and build the *nested hourglass networks* together with the stacked hourglass architecture.

2. Related Work

Human Pose Estimation Articulated human poses were usually modeled by combination of unary term and graph models, *e.g.*, mixture of body parts [44, 8] or pictorial structures [29]. Recently, significant progresses have been achieved by introducing ConvNets for learning better feature representation [38, 37, 36, 8, 42, 39, 31, 28]. For example, Chen and Yuille [8] introduced the ConvNet to learn both the unary and the pairwise term of a tree-structured graphical model. Tompson *et al.* [36] used multiple branches of ConvNets to fuse the features from an image pyramid, and used a Markov Random Field (MRF) for post-processing. Convolutional Pose Machine [39] incorporated the inference of the spatial correlations among body parts within the ConvNets. State-of-the-art performance is achieved by the stacked hourglass network [28] and its variant [5], which use repeated pooling down and upsampling process to learn the spatial distribution. Our approach is complementary to previous approaches by incorporating diverse image dependent multi-context representation to guide the human pose estimation.

Multiple Contextual Information The contextual information is generally referred to as regions surrounding the target locations [11, 13, 33], object-scene relationships [20, 18, 12], and object-object interactions [43]. It has been proved efficient in vision tasks as object classification [25] and detection [48, 11, 12]. Recent works modeled contextual information by concatenating multi-scale features [15, 14], or by gated functions to control the mutual influence of different contexts [49]. The contextual regions, however, are manually defined as rectangles without considering the objects appearance. In this work, we adopt visual attention mechanism to focus on regions which are image dependent and adaptiving for multi-context modeling. Our

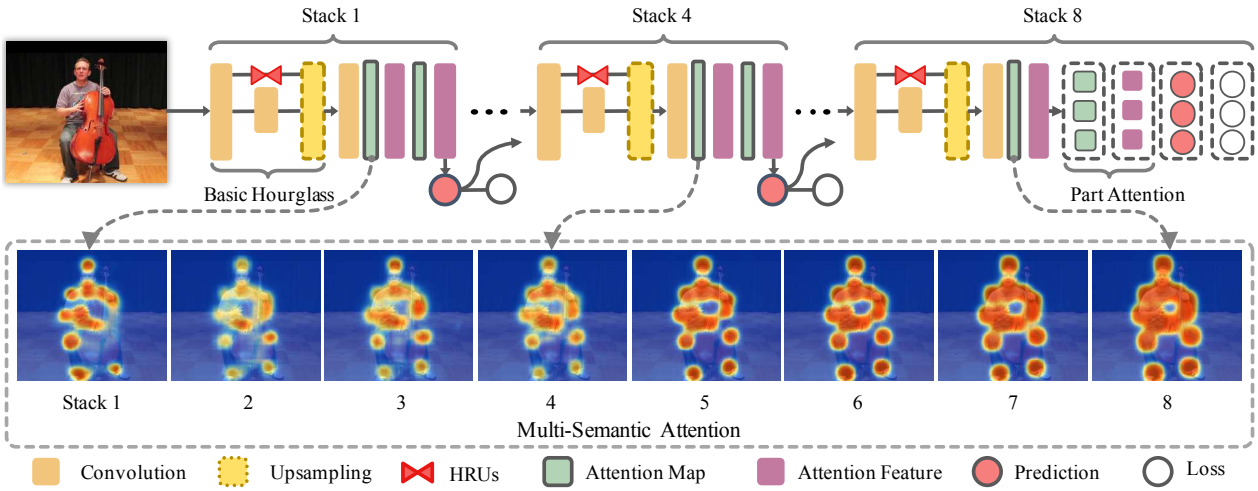

Figure 2. **Framework**. The basic structure is an 8-stack hourglass network. In each stack of hourglass, we generate *multi-resolution* attention maps. We also apply *multi-semantic* attention map to each hourglass as shown in stack 1 to stack 8. *Hierarchical Attention Mechanism* for zooming in on local parts is applied in stack 5 to stack 8.

approach increases the diversity of contexts.

Visual Attention Mechanism Since the visual attention model is computationally efficient and is effective in understanding images, it has achieved great success in various tasks such as machine translation [3], object recognition [2, 17, 6, 40], image captioning [47, 41], image question answering [46], and saliency detection [26]. Existing approaches usually adopted recurrent neural networks to generate the attention map for an image region at each step, and combined information from different steps overtime to make the final decision [3, 2, 26]. To the best of our knowledge, our work is the first to use attention models for human pose estimation. In addition, our design of the holistic attention map and the part attention map in learning attention in hierarchical order and the modeling of attention from different context and resolution are not investigated in these works.

3. Framework

An overview of our framework is illustrated in Fig. 2. In this section, we briefly introduce the nested hourglass architecture, and the implementation of the multi-context attention model, including the multi-semantics, multi-resolution, and hierarchical holistic-part attention model. The generated attention maps are then used to reweight the features for automatically inferring the regions of interest.

Baseline Network We adopt an 8-stack hourglass network [28] as the baseline network. It allows for repeated bottom-up, top-down inference across scales with intermediate supervision at the end of each stack. In experiments, the input images are 256×256, and the output heatmaps are $P \times 64 \times 64$, where P is the number of body parts. We follow previous work [36, 39, 28] to use the Mean Squared Error as the loss function.

Nested Hourglass Networks We replace the residual units, which are along the side branches for combining features across multiple resolutions, by the proposed micro hourglass residual units (HRUs), and obtain a *nested hourglass network* , as illustrated in Fig. 3. With this architecture, we enrich the information received by the output of each building block, which makes the whole framework more robust to scale change. Details of HRUs are described in Section 4.

Multi-Resolution Attention Within each hourglass, the multi-resolution attention maps Φ_r are generated from features of different scales, where r is the size of the features, as shown in Fig. 5. Attention maps are then combined to generate the refined features, which are further used to generate refined attention maps and further refined features, as shown in Fig. 4.

Multi-Semantics Attention Different stacks are with different semantics: lower stacks focus on local appearance, while higher stacks encode global representations. Hence attention maps generated from different stacks also encode various semantic meanings. As shown in Fig. 2, compare the *left knee* in Stack 1 with 8, we can see that deeper stacks with global representations are able to recover occlusions.

Hierarchical Attention Mechanism In the lower stacks, *i.e.*, stack 1 to stack 4, we use two holistic attention maps $\mathbf{h}_1^{\text{att}}$ and $\mathbf{h}_2^{\text{att}}$ to encode configurations of the whole human body. In the higher stacks, *i.e.*, the 5th to the 8th stack, we design a hierarchical coarse-to-fine attention scheme to zoom into local parts.

4. Nested Hourglass Networks

In this section, we provide a detailed description of the proposed hourglass residual units (HRUs). We also provide comprehensive analysis of the receptive field.

4.1. Hourglass Residual Units

Let us first briefly recall Residual networks [19]. Deep residual networks achieve compelling accuracy by an extremely deep stacks of "Residual Units", which can be expressed as follows,

$$\mathbf{x}_{n+1} = h(\mathbf{x}_n) + \mathcal{F}(\mathbf{x}_n, \mathbf{W}_n^{\mathcal{F}}), \tag{1}$$

where $\mathbf{x}_n$ and $\mathbf{x}_{n+1}$ are the input and output of the n-th unit, and $\mathcal{F}$ is the stacked convolution, batch normalization, and ReLU nonlinearity. In [19], $h(\mathbf{x}_n) = \mathbf{x}_n$ is the identity mapping.

In this paper, we focus on human pose estimation, in which larger contextual regions are proved to be important for locating local body parts [39, 28]. The contextual region of a neuron is its corresponding receptive field. In this work, we propose to extend the original residual units by a micro hourglass branch. The resulted hourglass residual units (HRUs) have larger receptive field while preserve local details, as shown in Fig. 3. We use this module in the stacked hourglass networks. This architecture is referred to as "nested hourglass networks" because the hourglass structure is used at both the macro and micro levels.

The mathematical formulation of our proposed HRUs is as follows:

$$\mathbf{x}_{n+1} = \mathbf{x}_n + \mathcal{F}(\mathbf{x}_n, \mathbf{W}_n^{\mathcal{F}}) + \mathcal{P}(\mathbf{x}_n, \mathbf{W}_n^{\mathcal{P}}). \tag{2}$$

Each HRU consists of three branches. Branch (A), *i.e.* $\mathbf{x}_n$ in (2), is the identity mapping. Hence, the property of ResNet in handling vanishing gradient is preserved in the HRUs. Branch (B), *i.e.* $\mathcal{F}(\mathbf{x}_n, \mathbf{W}_n^{\mathcal{F}})$ in (2), is the residual block like the ResNet in (1). Branch (C), *i.e.* $\mathcal{P}(\mathbf{x}_n, \mathbf{W}_n^{\mathcal{P}})$ in (2), is our new design, which is a stack of a 2×2 max-pooling, two 3×3 convolutions followed by ReLU nonlinearity, and an upsampling operation.

4.2. Analysis of Receptive Field of HRU

The identity mapping in branch (A) has receptive size of one. The residual block in branch (B) is a stack of convolutions ($\text{Conv}_{1\times1} + \text{Conv}_{3\times3} + \text{Conv}_{1\times1}$). Hence, the neuron in the output feature corresponds to a 3×3 region of the input in this HRU. Branch (C) is our added branch. The structure of this branch is $\text{Pool}_{2\times2} + \text{Conv}_{3\times3} + \text{Conv}_{3\times3} + \text{Deconv}_{2\times2}$. Due to max-pooling, the resolution for convolution in this branch is half of that in branches (A) and (B), and each neuron in the output feature map corresponds to a 10×10 region of the input, which is about 3 times the receptive field size of the residual block in branch (B). These three branches, with different receptive fields and resolutions, are added together as the output of the HRU. Therefore, the HRU unit increases the receptive field size by including the branch (C) while preserves the high-resolution information by using branches (A) and (B).

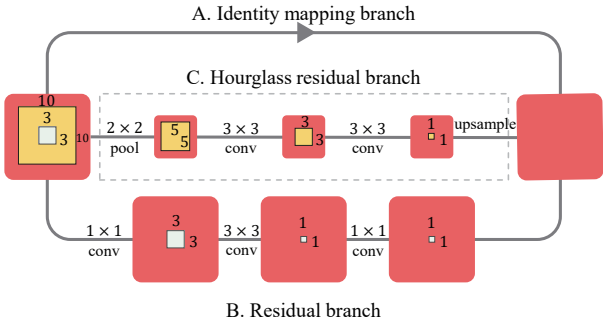

Figure 3. An illustration of the *hourglass residual unit*. It consists of three branches: (A) identity mapping, (B) residual branch, and (C) hourglass residual branch. The receptive fields with respect to the input are 3×3 and 10×10 for the conventional residual branch and the hourglass residual branch, respectively.

5. Attention Mechanism

We shall first briefly introduce the conventional soft attention mechanism, and then describe our proposed multi-context framework.

5.1. Conventional Attention

Denote convolutional features by $\mathbf{f}$. The first step in obtaining soft attention is to generate the summarized feature map as follows:

$$\mathbf{s} = g(\mathbf{W}^a * \mathbf{f} + \mathbf{b}), \tag{3}$$

where $*$ denotes convolution, $\mathbf{W}^a$ denotes the convolution filters, and g is the nonlinear activation function. $\mathbf{s} \in \mathbb{R}^{H \times W}$ summarizes information of all channels in $\mathbf{f}$.

Denote $\mathbf{s}(l)$ as the feature at location l in the feature map $\mathbf{s}$, where $l = (x, y)$, x is the horizontal location and y is the vertical location. The Softmax operation is applied to $\mathbf{s}$ spatially as follows:

$$\Phi(l) = \frac{e^{\mathbf{s}(l)}}{\sum_{l' \in \mathbb{L}} e^{\mathbf{s}(l')}}, \tag{4}$$

where $\mathbb{L} = \{(x, y) | x = 1, \ldots, W, y = 1, \ldots, H\}$. Φ is the attention map, where $\sum_{l \in \mathbb{L}} \Phi(l) = 1$. Then the attention map is applied to the feature $\mathbf{f}$,

$$\mathbf{h}^{\text{att}} = \Phi \star \mathbf{f}, \qquad \text{where } \mathbf{h}^{\text{att}}(c) = \mathbf{f}(c) \circ \Phi, \tag{5}$$

where c is the index for feature channel. We use $\star$ to represent the channel-wise Hadamard matrix product operation. $\mathbf{h}^{\text{att}}$ is the refined feature map, which is the feature reweighted by the attention map, and has the same size as $\mathbf{f}$.

5.2. Our Multi-Context Attention Model

Our framework makes the following three modifications to the attention model. First, we replace the global Softmax in 4 with a CRF to taking local pattern correlations into consideration. Global spatial Softmax normalizes the whole image based on a constant factor, which ignores the local neighboring spatial correlations. But we want attention maps to drive the network to concentrate on the complex human body configurations. More details are in Section 5.2.1. Second, we generate attention maps based on

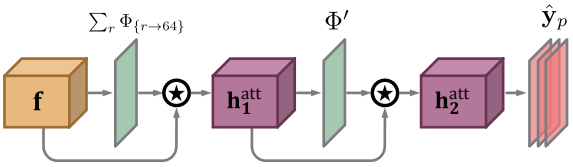

Figure 4. An illustration of the attention scheme.

features of different resolutions to make the model more robust, as illustrated in Section 5.2.2. Then multi-semantics attention is obtained by generating attention maps for each stack of the hourglass, as described in Section 5.2.3. Finally, a hierarchical coarse to fine(*i.e.* fully body to parts) attention scheme is used, to zoom into local part regions for more precise localization, which is introduced in Section 5.2.4. The whole framework is differentiable and trained end-to-end with random initialization. An illustration of our attention scheme is shown in Fig. 4.

5.2.1 Spatial CRF Model

In this work, we use Conditional Random Fields (CRFs) to model the spatial correlation. To make them differentiable, we use the mean-field approximation approach to recursively learn the spatial correlation kernel [50, 24].

The attention map is modeled as a two-class problem. Denote $y_l = \{0, 1\}$ as the attention label at the i-th location. In the CRF model, the energy of a label assignment $\mathbf{y} = \{y_l | l \in \mathbb{L}\}$ is as follows:

$$E(\mathbf{z}) = \sum_l y_l \psi_u(l) + \sum_{l,k} y_l w_{l,k} y_k, \tag{6}$$

where $\psi(y_l) = g(\mathbf{h}, l)$ is the unary term that measures the inverse likelihood (and therefore, the cost) of the position l taking the attention label $y_l = 1$. $w_{l,k}$ is the weight for compatibility between y_l and y_k. Given the image $\mathbf{I}$, the probability of the label assignment $\mathbf{y}$ is $P(\mathbf{y}|\mathbf{I}) = \frac{1}{Z}\exp(-E(\mathbf{y}|\mathbf{I}))$, where Z is the partition function. The probability for $y_l = 1$ is obtained iteratively using the mean-field approximation as follows:

$$\Phi(y_l = 1)_t = \sigma\left(\psi_u(l) + \sum_k w_{l,k}\Phi(y_k = 1)_{t-1}\right), \tag{7}$$

where $\sigma(a) = 1/(1 + \exp(-a))$ is the sigmoid function. $\psi_u(l)$ is obtained by convolution from features $\mathbf{h}$. $\sum_k w_{l,k}\Phi(y_j = 1)$ is implemented by convolving the estimated attention map Φ_{t-1} at the stage $t-1$ with the filters. Initially, $\Phi(y_i = 1)_1 = \sigma(\psi_u(i))$.

In summary, the attention map Φ_t at the stage t can be formulated as follows:

$$\Phi_t = \mathcal{M}(\mathbf{s}, \mathbf{W}^k) = \begin{cases} \sigma(\mathbf{W}^k * \mathbf{s}) & t = 0, \\ \sigma(\mathbf{W}^k * \Phi_{t-1}) & t = 1, 2, 3, \end{cases} \tag{8}$$

where $\mathcal{M}$ denotes a sequence of weights-sharing convolutions for the mean field approximation, $\mathbf{W}^k$ denotes the spatial correlation kernel. The $\mathbf{W}^k$ is shared across different time steps. In our network, we use three steps of recursive convolution.

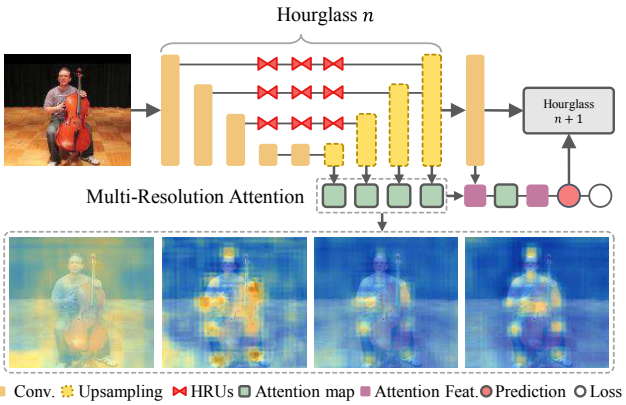

Figure 5. The multi-resolution attention scheme within an hourglass. In each stack of hourglass, we generate multi-resolution attention maps from features with different resolutions. These maps are summed into a single attention map, which applies to features $\mathbf{f}$ to generate the refined feature $\mathbf{h}_1^{\text{att}}$.

5.2.2 Multi-Resolution Attention

As shown in Fig. 5, the upsampling process generates features of different size r, *i.e.* $\mathbf{f}_r$ for $r = 8, 16, 32$ and 64. $\mathbf{s}_r$ is used to generate the attention map Φ_r using the procedure in (8). The attention map Φ_r is upsampled to size 64, which is denoted by $\Phi_{\{r\to 64\}}$. These attention maps correspond to different resolutions. As shown in Fig. 5 (I), $\Phi_{\{8\to 64\}}$, which has lower resolution, and highlights the whole configure of human body. Φ_{64}, which is generated with higher resolution, focusing on local body parts.

All up-sampled attention maps are summed up and then applied to the feature $\mathbf{f}$,

$$\mathbf{h}_1^{\text{att}} = \mathbf{f} \star \left(\sum_{r=8,16,32,64} \Phi_{\{r\to 64\}}\right), \tag{9}$$

where the feature $\mathbf{f}$ is the output of the last layer in an hourglass stack as shown in Fig. 5. The operation $\star$ is illustrated in Eq. (5).

The conventional way of using an attention map is to directly apply it to the feature which generates it. However, the features refined by attention map usually have large amount of values close to zero, and so a stack of many refined features makes the back-propagation difficult. To utilize information from multi-resolution features without sacrificing training efficiency, we generate attention maps from features with various resolutions, and apply them to the later features.

In addition to the multi-resolution attention, a refined attention map Φ' and its corresponding refined feature $\mathbf{h}_2^{\text{att}}$ are generated from $\mathbf{h}_1^{\text{att}}$,

$$\mathbf{h}_2^{\text{att}} = \mathbf{h}_1^{\text{att}} \star \Phi' = \mathbf{h}_1^{\text{att}} \star \mathcal{M}(\mathbf{h}_1^{\text{att}}, \mathbf{w}). \tag{10}$$

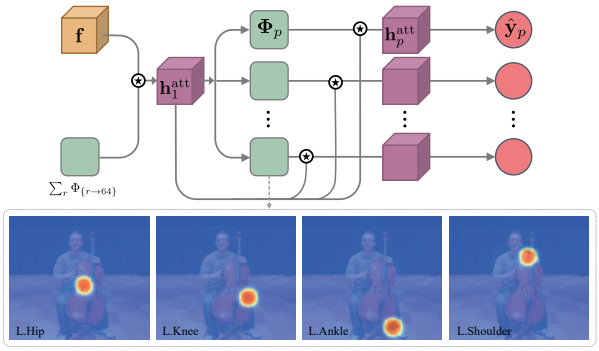

Figure 6. Coarse-to-fine part attention model and the visualization of examplar part attention maps.

5.2.3 Multi-Semantics Attention

The above procedure is repeated over stacks of hourglass to generate attention maps with multiple semantic meanings. Samples of Φ' are shown in Fig. 2 from stack 1 to 8. The attention maps at shallower hourglass stacks capture more local information. For deeper hourglass stacks, the global information about the whole person is captured, which is more robust to occlusion.

5.2.4 Hierarchical Holistic-Part Attention

In the 4th to 8th stacks of hourglass structure, we use the the refined feature $\mathbf{h}_1^{\text{att}}$ in Eq. (9) to generate the part attention maps as follows:

$$\begin{aligned} \mathbf{s}_p &= g(\mathbf{W}_p^a * \mathbf{h}_1^{\text{att}} + \mathbf{b}), \\ \Phi_p &= \mathcal{M}(\mathbf{s}_p, \mathbf{W}_p^k), \end{aligned} \tag{11}$$

where $p \in \{1, \cdots, P\}$, $\mathbf{W}_p^a$ denotes the parameters for obtaining the summarization map $\mathbf{s}_p$ of part p, $\mathbf{W}_p^k$ denotes the spatial correlation modeling for part p. The part attention map Φ_p is combined with the refined feature map $\mathbf{h}_1^{\text{att}}$ to obtain the refined feature map for part p as follows:

$$\mathbf{h}_p^{\text{att}} = \mathbf{h}_1^{\text{att}} \star \Phi_p. \tag{12}$$

The heatmap predication for the pth body joint is based on the refined features $\mathbf{h}_p^{\text{att}}$,

$$\hat{\mathbf{y}}_p = \mathbf{w}_p^{\text{cls}} * \mathbf{h}_p^{\text{att}}, \tag{13}$$

where $\hat{\mathbf{y}}_p$ is the heatmap for the pth part, $\mathbf{w}_p^{cls}$ is the classifier. In this way, we guarantee that the attention map Φ_p is specific for the body joint p. Some qualitative results of part attention maps are shown in Fig. 6.

6. Training the model

Each stack in the hourglass produces the estimated heatmaps for the body joints. We adopt the loss function in [28] for learning the model. For each stack, the Mean Squared Error (MSE) loss is computed by

$$L = \sum_{p=1}^{P} \sum_{l \in \mathbb{L}} \|\hat{\mathbf{y}}_p(l) - \mathbf{y}_p(l)\|_2^2 \tag{14}$$

where p denotes the pth body part, l denotes the lth location. $\hat{\mathbf{y}}_p$ denotes the predicted heatmap for part p, and $\mathbf{y}_p$ the corresponding ground-truth heatmap generated by a 2-D Gaussian centered on the body part location.

The attention maps help to drive the network to focus on hard negative samples. After several stages of training, the attention maps fire on human body region, where the true positive samples are highlighted by attention maps. The refined features are used for learning classifiers for the regions with human body, with easy background regions removed at the feature level by the learned attention maps. Consequentially, for part attention maps, the classifiers focus on classifying each body joint based on well defined human body regions, without considering the background.

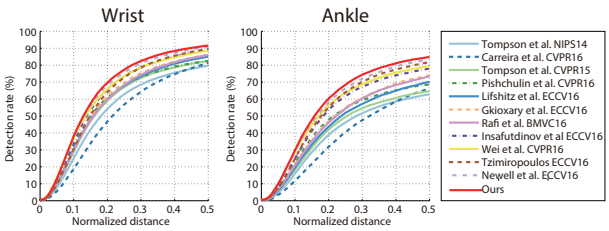

Figure 7. Comparisons of PCKh curves on the MPII Human Pose test set on the most challenging body joints, *i.e.*, *wrist* and *ankle*.

7. Experiments

Dataset We evaluate the proposed method on two widely used benchmarks, MPII Human Pose [1] and extended Leeds Sports Poses (LSP) [23]. The MPII Human Pose dataset includes about 25k images with 40k annotated poses. The images were collected from YouTube videos covering daily human activities with highly articulated human poses. The LSP dataset consists of 11k training images and 1k testing images from sports activities.

Data Augmentation During training, we crop the images with the target human centered at the images with roughly the same scale, and warp the image patch to the size 256×256. Then we randomly rotate ($\pm 30°$) and flip the images. We also perform random rescaling (0.75 to 1.25) and color jittering to make the model more robust to scale and illumination change. During testing, we follow the standard routine to crop image patches with the given rough position and the scale of the test human for MPII dataset. For the LSP dataset, we simply use the image size as the rough scale, and the image center as the rough position of the target human to crop the image patches. All the experimental results are produced from the original and flipped image pyramids with 6 scales.

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
Pishchulin *et al.* [30]	74.3	49.0	40.8	34.1	36.5	34.4	35.2	44.1
Tompson *et al.* [37]	95.8	90.3	80.5	74.3	77.6	69.7	62.8	79.6
Carreira *et al.* [7]	95.7	91.7	81.7	72.4	82.8	73.2	66.4	81.3
Tompson *et al.* [36]	96.1	91.9	83.9	77.8	80.9	72.3	64.8	82.0
Hu&Ramanan [21]	95.0	91.6	83.0	76.6	81.9	74.5	69.5	82.4
Pishchulin *et al.* [31]	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4
Lifshitz *et al.* [27]	97.8	93.3	85.7	80.4	85.3	76.6	70.2	85.0
Gkioxary *et al.* [16]	96.2	93.1	86.7	82.1	85.2	81.4	74.1	86.1
Rafi *et al.* [32]	97.2	93.9	86.4	81.3	86.8	80.6	73.4	86.3
Insafutdinov *et al.* [22]	96.8	95.2	89.3	84.4	88.4	83.4	78.0	88.5
Wei *et al.* [39]	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Bulat&Tzimiropoulos [5]	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7
Newell *et al.* [28]	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9
Ours	**98.5**	**96.3**	**91.9**	**88.1**	**90.6**	**88.0**	**85.0**	**91.5**

Table 1. Comparisons of PCKh@0.5 score on the MPII test set.

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
Belagiannis&Zisserman [4]	95.2	89.0	81.5	77.0	83.7	87.0	82.8	85.2
Lifshitz *et al.* [27]	96.8	89.0	82.7	79.1	90.9	86.0	82.5	86.7
Pishchulin *et al.* [31]	97.0	91.0	83.8	78.1	91.0	86.7	82.0	87.1
Insafutdinov *et al.* [22]	97.4	92.7	87.5	84.4	91.5	89.9	87.2	90.1
Wei *et al.* [39]	97.8	92.5	87.0	83.9	91.5	90.8	89.9	90.5
Bulat&Tzimiropoulos [5]	97.2	92.1	88.1	85.2	92.2	91.4	88.7	90.7
Ours	**98.1**	**93.7**	**89.3**	**86.9**	**93.4**	**94.0**	**92.5**	**92.6**

Table 2. Comparisons of PCK@0.2 score on the LSP dataset.

Experiment Settings We train our model with Torch7 [10] using the initial learning rate of 2.5×10^{-4}. The parameters are optimized by RMSprop [35] algorithm. We train the model on the MPII dataset for 130 epochs and the LSP dataset for 60 epochs. We adopt the validation split for the MPII dataset used in [36] to monitor the training process.

7.1. Results

We use the Percentage Correct Keypoints (PCK) [45] metric for comparisons on the LSP dataset, and the PCKh measure [1], where the error tolerance is normalized with respect to head size, for comparisons on the MPII Human Pose dataset.

MPII Human Pose Table 1 reports the comparison of the PCKh performance of our method and previous state-of-the-art at a normalized distance of 0.5. Our method achieves state of the art 91.5% PCKh scores. In particular, for the most challenging body parts, *e.g.*, *wrist* and *ankle*, our method achieves 1.0% and 1.4% improvement compared with the closed competitor respectively, as shown in Fig. 7.

Leeds Sports Pose We train our model by adding the MPII training set to the extended LSP training set with *person-centric* annotations, which is a standard routine [39, 22, 31, 27, 4]. Table 2 reports the PCK at threshold of 0.2. Our approach outperforms the state-of-the-art across all the body joints, and obtains 1.9% improvement in average.

7.2. Component Analysis

To investigate the efficacy of the proposed multi-context attention mechanism and the hourglass residual unit, we conduct ablation experiments on the validation set [36] of the MPII Human Pose dataset. We use an 8-stack hourglass network [28] as our baseline model if not specified. The overall result is shown in Fig. 8. Based on the baseline network (BL), we analyze each proposed component, *i.e.*, the Multi-Semantics attention model (MS), Hourglass Residual Units (HRUs), Multi-Resolution attention model (MR), and the Hierarchical Part attention model (HP), by comparing the PCKh score.

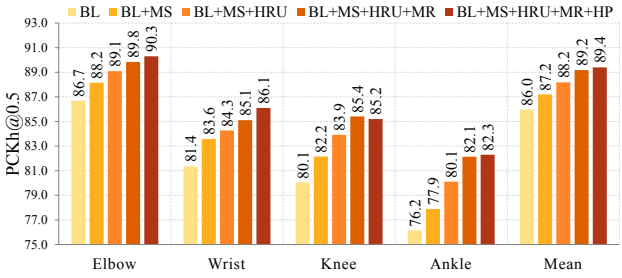

Figure 8. **Component analysis**. PCKh scores at threshold of 0.5 on the MPII validation set.

Multi-Semantics Attention We first evaluate the multi-semantics attention model. By adding holistic attention model at the end of each stack of hourglass ("BL+MS"), we get an 87.2% PCKh score, which is a 1.2% improvement compared to the baseline model.

Hourglass Residual Unit To explore the effect of the residual pooling unit, we further use the HRUs to replace the original residual units when combining features from different resolutions ("BL+MS+HRU"), as illustrated in Fig. 2. The addition of hourglass residual unit results in a further 1% improvement. As discussed in [28], improvements cannot be easily obtained by simply stacking more than eight hourglass modules. We provide a way to increase the model capacity effectively.

Multi-Resolution Attention By generating attention maps from features with multiple resolutions ("BL+MS+HRU+MR"), our method obtains a further 1% improvement.

Hierarchical Attention We also show the improvement brought by the hierarchical holistic-local attention model. We replace the refined holistic attention map by a set of part attention maps from stack four to eight, and obtain the highest mean PCKh score 89.4%. We observe the improvements are mostly brought by the refined localization of body parts. In some cases, the part attention model could even correct the double counting problem, as demonstrated in Fig. 1 (c).

Softmax vs. CRF Finally, we compare the proposed CRF spatial attention model with the conventional Softmax attention model based on a 2-stack hourglass network. We compare the accuracy rates, *i.e.*, PCKh at 0.5, on the validation set as training progresses in Fig. 10. The CRF attention model converges much faster and achieves higher

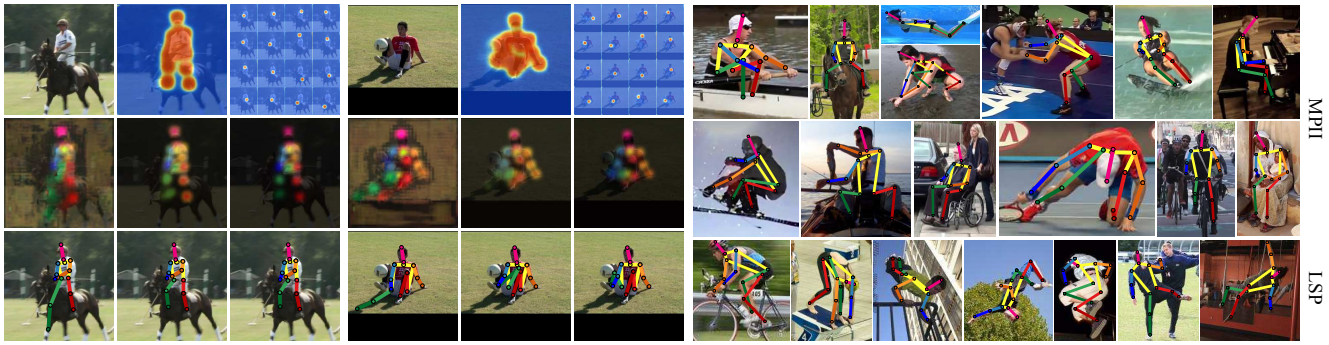

Figure 9. **Qualitative evaluation**. (a-b) 1st row to 3rd row: 2 input images, 4 attention maps, 6 heatmaps, and 6 predicted poses. (c) Examples of estimated poses on the MPII test set and the LSP test set (Best viewed in electronic form with 4× zoom in).

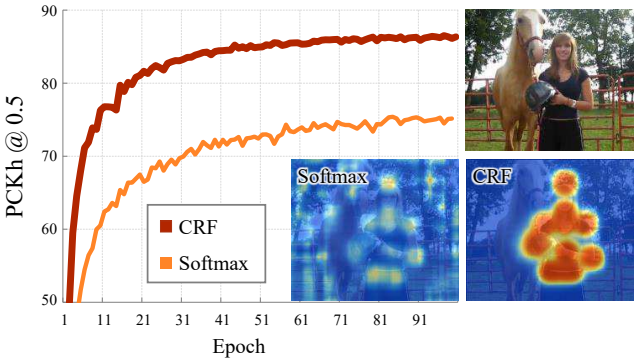

Figure 10. PCKh@0.5 on the MPII validation set across training.

validation accuracy than the Softmax attention model. We visualize the attention maps generated by these two models, and observe that CRF attention model generates much more cleaner attention maps compared with Softmax attention model due to its better ability to model spatial correlations among body parts.

7.3. Qualitative Results

To gain insights on how attention works, we compare the baseline model with the proposed model by visualizing the attention maps, the score maps, and the estimated poses, as demonstrated in Fig. 9 (a-b). We observe the baseline model may have difficulty in distinguishing objects with similar appearance with limbs (*e.g.*, the horse leg in Fig. 9 (a)), and the heavy shadow with ambiguous shape (Fig. 9 (b)). So the holistic attention maps would be great help for removing cluttered background and reducing ambiguity. For part attention maps, besides providing more precise localization for the body parts, they could even help reduce the double counting problem. For example, the left and right *ankle* can be distinguished by incorporating the part attention maps.

Fig. 9 (c) demonstrates the poses predicted by our methods on the MPII test set and the LSP test set. Our method is robust to extremely difficult cases, *e.g.*, rare poses, cluttered background, and foreshortening. However, as shown

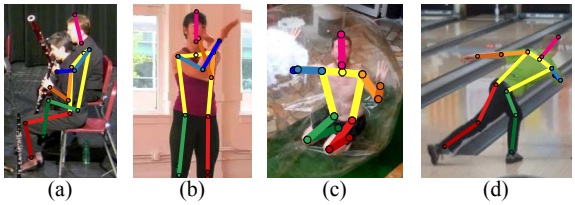

Figure 11. Failure cases caused by (a) overlapping people, (b) twisted limbs, (c) illumination, and (d) left/right confusion.

in Fig. 11, our method may fail in some cases which are also difficult for human eyes, *i.e.* (a) heavy occlusion and ambiguity, (b) twisted limbs, (c) significant illumination change, and (d) left/right body confusion caused by clothing/lighting.

8. Conclusion

This paper has proposed incorporating multi-context attention and ConvNets into an end-to-end framework. We use visual attention to guide context modeling. Hence our framework has large diversity in contextual regions. Instead of using global Softmax, we introduce CRF for spatial correlation modeling. We build multi-context attention model along three components, *i.e.*, multi-resolution, multi-semantics, and hierarchical holistic-part attention scheme. Additionally, an hourglass residual unit was proposed to enrich the expressive power of conventional residual unit. The proposed multi-context attention and the HRUs are general, and would help other vision tasks.

Acknowledgment: This work is supported by SenseTime Group Limited, the General Research Fund sponsored by the Research Grants Council of Hong Kong (Project Nos. CUHK14213616, CUHK14206114, CUHK14205615, CUHK419 412, CUHK14203015, and CUHK14207814), the Hong Kong Innovation and Technology Support Programme (No.ITS/121/15FX), National Natural Science Foundation of China (Nos. 61371192, 61301269), PhD programs foundation of China (No. 20130185120039), and ONR N00014-15-1-2356.

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 2, 6, 7

[2] J. Ba, V. Mnih, and K. Kavukcuoglu. Multiple object recognition with visual attention. In *ICLR*, 2015. 3

[3] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015. 3

[4] V. Belagiannis and A. Zisserman. Recurrent human pose estimation. *arXiv preprint arXiv:1605.02914*, 2016. 7

[5] A. Bulat and G. Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *ECCV*, 2016. 2, 7

[6] C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, W. Xu, et al. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *ICCV*, 2015. 3

[7] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human pose estimation with iterative error feedback. In *CVPR*, 2016. 7

[8] X. Chen and A. L. Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *NIPS*, 2014. 2

[9] X. Chu, W. Ouyang, H. Li, and X. Wang. Structured feature learning for pose estimation. In *CVPR*, 2016. 1

[10] R. Collobert, K. Kavukcuoglu, and C. Farabet. Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*, 2011. 7

[11] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 2

[12] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert. An empirical study of context in object detection. In *CVPR*, 2009. 2

[13] X. Fan, K. Zheng, Y. Lin, and S. Wang. Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. In *CVPR*, 2015. 2

[14] S. Gidaris and N. Komodakis. Object detection via a multi-region and semantic segmentation-aware cnn model. In *ICCV*, 2015. 1, 2

[15] R. Girshick, F. Iandola, T. Darrell, and J. Malik. Deformable part models are convolutional neural networks. In *CVPR*, 2015. 1, 2

[16] G. Gkioxari, A. Toshev, and N. Jaitly. Chained predictions using convolutional neural networks. In *ECCV*, 2016. 7

[17] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, 2015. 3

[18] A. Gupta, T. Chen, F. Chen, D. Kimber, and L. S. Davis. Context and observation driven latent variable model for human pose estimation. In *CVPR*, 2008. 2

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2, 4

[20] G. Heitz and D. Koller. Learning spatial context: Using stuff to find things. In *ECCV*. 2008. 2

[21] P. Hu and D. Ramanan. Bottom-up and top-down reasoning with hierarchical rectified gaussians. In *CVPR*, 2016. 7

[22] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *ECCV*, 2016. 7

[23] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. 2, 6

[24] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011. 5

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1, 2

[26] J. Kuen, Z. Wang, and G. Wang. Recurrent attentional networks for saliency detection. In *CVPR*, 2016. 2, 3

[27] I. Lifshitz, E. Fetaya, and S. Ullman. Human pose estimation using deep consensus voting. In *ECCV*, 2016. 7

[28] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 1, 2, 3, 4, 6, 7

[29] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Poselet conditioned pictorial structures. In *CVPR*, 2013. 2

[30] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Strong appearance and expressive spatial models for human pose estimation. In *ICCV*, 2013. 7

[31] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *CVPR*, June 2016. 2, 7

[32] U. Rafi, J. Gall, and B. Leibe. An efficient convolutional network for human pose estimation. In *ECCV*, 2016. 7

[33] V. Ramakrishna, D. Munoz, M. Hebert, J. A. Bagnell, and Y. Sheikh. Pose machines: Articulated pose estimation via inference machines. In *ECCV*. 2014. 1, 2

[34] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015. 2

[35] T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 4(2), 2012. 7

[36] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler. Efficient object localization using convolutional networks. In *CVPR*, 2015. 1, 2, 3, 7

[37] J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS*, 2014. 1, 2, 7

[38] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, 2014. 1, 2

[39] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016. 1, 2, 3, 4, 7

[40] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *CVPR*, 2015. 3

[41] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2(3):5, 2015. 2, 3

[42] W. Yang, W. Ouyang, H. Li, and X. Wang. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. CVPR, 2016. 2

[43] Y. Yang, S. Baker, A. Kannan, and D. Ramanan. Recognizing proxemics in personal photos. In *CVPR*, 2012. 2

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

[45] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *PAMI*, 35(12):2878–2890, 2013. 7

[46] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. *arXiv preprint arXiv:1511.02274*, 2015. 2, 3

[47] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. *arXiv preprint arXiv:1603.03925*, 2016. 2, 3

[48] X. Zeng, W. Ouyang, and X. Wang. Multi-stage contextual deep learning for pedestrian detection. In *ICCV*, 2013. 2

[49] X. Zeng, W. Ouyang, J. Yan, H. Li, T. Xiao, K. Wang, Y. Liu, Y. Zhou, B. Yang, Z. Wang, et al. Crafting gbd-net for object detection. *arXiv preprint arXiv:1610.02579*, 2016. 1, 2

[50] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. Torr. Conditional random fields as recurrent neural networks. In *ICCV*, 2015. 5