Multi-scale Adaptive Structure Network for Human Pose Estimation from Color Images

Wenlin Zhuang[1], Cong Peng[2], Siyu Xia[1], and Yangang, Wang[1]*

[1] School of Automation, Southeast University, Nanjing, China
[2] School of Automation, Nanjing University of Aeronautics and Astronautics, Nanjing, China

Abstract. Human pose estimation is formulated as a joint heatmap regression problem in the deep learning based methods. Existing convolutional neural networks usually adopt fixed kernel size for generating joint heatmaps without regard to the size of human shapes. In this paper, we propose a novel method to address this issue by adapting the kernel size of joint heatmaps to the human scale of the input images in the training stage. We present a normalization strategy of how to perform the adaption between the kernel size and human scale. Beyond that, we introduce a novel limb region representation to learn the human pose structural information. Both the adaptive joint heatmaps as well as the limb region representation are combined together to construct a novel neural network, which is named **Multi-scale Adaptive Structure Network (MASN)**. The effectiveness of the proposed network is evaluated on two widely used human pose estimation benchmarks. The experiments demonstrate that our approach could obtain the state-of-the-art results and outperform the most existing methods over all the body parts.

Keywords: multi-scale · adaptive heatmaps · human pose estimation

1 Introduction

Human pose estimation, which is defined as the problem of localization of human skeleton joints, is a vital yet challenging task in the field of computer vision. It has many important applications such as human-machine interactions, intelligent manufacturing, autonomous driving and etc. In recent years, human pose estimation from single RGB image has gained significant improvements by Convolutional Neural Networks (ConvNets) [1–6]. However, achieving accurate human pose from color images is still very difficult due to variant appearances, strong articulations, heavy occlusions and etc.

The main stream of work for human pose estimation from color images with deep learning has been motivated by formulating it as a joint regression problem, where joint coordinates [1] or joint heatmaps [2] are the most widely used human pose representation. Since joint heatmaps are robust to the data noise and have

* Corresponding author: Yangang Wang. E-mail: ygwangthu@gmail.com

better performance [2], they attract more and more attentions from researchers in this field. Typically, a joint heatmap is described as a 2D map with a circle at the position of the joint. The values in the circle are computed from a 2D Gaussian probability density function. The variance of the density function determines the influence range of the circle in the joint heatmap. In this paper, the variance is named as the **kernel size** of the joint heatmap.

Existing methods [7, 8] usually adopt fixed kernel size without considering the the size of human shapes in the input image, which is named as **human scale** in this paper. However, we argue that the human scale is very important for regressing the accurate joint heatmaps. Suppose there is a very small person in a given input image, fixed kernel size might generate joint heatmaps which cover the whole pixels of the human. This scenario is not what we expect and we want the neural network would have available receptive fields according to the image content. In other words, the kernel size of joint heatmaps should be appropriate to the human scale in the given input images. We want the designed neural network could have the abilities that large person has large kernel size and small person has small kernel size. The 'large' and 'small' are relative to the coverage ratio of human in the same size of rescaled input image. More importantly, the joint heatmaps for different human scales should be normalized into a unified framework while performing the end-to-end training with the convolutional neural networks.

It is noted that several previous works attempt to address the human scales in the methods of deep learning based human pose estimation. Data augmentation with rescaling the input images [9] is the widely used strategy. Nevertheless, none of previous data augmentation with scaling strategies construct a concrete normalization between the human scale and human pose. Feature pyramid network [6] explicitly consider several different sizes of human shapes by constructing the pyramid structures of image features. However, they can only address limited scales (always 3 or 4) in the network structure. And their method does not consider the image normalization among the input dataset. Our key idea is to adapt the kernel size of joint heatmaps to the human scale of input images in the training stage. We deduct a mathematical normalization strategy (Sec. 3) to perform the adaption between the kernel size and human scale. We demonstrate that the normalized human scale as well as human pose could benefit with each other in an end-to-end neural network framework and has superior performance as described in Sec. 4.

In order to further improve the performance of human pose estimation, we also encode the pose structural information by referring the idea of part affinity [7]. Different from their directional representation of limb, we propose a simple yet effective limb region representation to capture the pose structural information. The limb region is composed of the pixels covered by the connecting line between two adjacent joints. We generate the limb region maps, where only the pixels in the limb region are filled with 1. Similar as the proposed adaptive joint heatmaps, the limb region maps are also associated with the human scales. Both the adaptive joint heatmaps as well as the limb region maps are combined together to construct a novel neural network, which is named as **Multi-scale Adaptive Structure Network (MASN)**. For exploiting the abilities of the proposed method, we use the Hourglass module [5] to construct our network architecture. The whole network structure is visualized in Fig. 1. We evaluate the proposed method on the MPII Human Pose dataset[9], and there is a clear average accuracy improvement over the original Hourglass network [5]. In addition, our model has fewer parameters, as shown in Table 1. In the following sections, we will introduce each individual components of the proposed method in details.

2 Related Works

Human Pose Estimation. Early human pose estimation usually adopt pictorial structure [10][11][12][13][14]. Due to the influence of human flexbility and occlusion, the application conditions are too harsh and these approachs are not robust enough. In recent years, pose estimation has been greatly developed under deep learning [15][1][16][3][2]. DeepPose [1] uses ConvNets to regress the coordinates of body joints. However, ConvNets cannot learn the mapping from image to location faultlessly.Therefore, the method of directly detecting heatmap appears, and its effect is better and more robust.Convolutional Pose Machine [4] is one of first adpot heatmap, which show ConvNets can learn image features and implicitly model long-range dependencies for different body parts. Hourglass [5], the most famous method in recent years, differs from other networks mainly in its more symmetric distribution of capacity between bottom-up processing (from high resolutions to low resolutions) and top-down processing (from low resolutions to high resolutions). Based on this, chu et al.[17] added attention to different resolutions of Hourglass.Yang et al. [6] made improvements to the basic module (Pyramid Residual Module) in Hourglass, and can obtain richer features at each scale. The detecting heatmap methods used above use the same kernel size of joint heatmaps. However, the shape of human body in two-dimensional images is different. For different human body shape, different kernel size should be selected. Our method uses an adaptive heatmap generation method to generate heatmaps of different kernel size according to different human body shape in color image.

Human Pose Structure. The human skeleton structure is the most critical information of human pose, which is crucial for inferring joints that are invisible under the influence of occlusion, illumination and etc. The pictorial structure [13] is based on human body structure using a visual description method to build a model of the human body structure. Chu et al. [18] used geometrical transform kernels in ConvNets to establish the dependencies between joints, and use a bi-directional tree to model the human upper body structure. Ke et al. [8] use Hourglass to learn body structure information to characterize body structure information using multiple joint heatmaps. These methods have different defects, and the structure of the human body is incomplete. Openpose [7] models the areas between the joints of the two connections, and uses the human skeleton

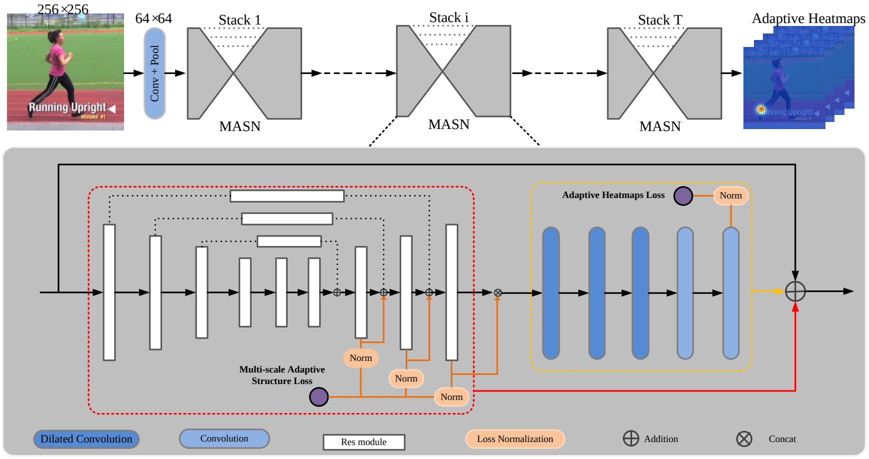


Fig. 1. Overview of our framework. **Top**: the Multi-scale Adaptive Structure Network(MASN), stacked T MASN sequentially. **Bottom**: details of each stack of MASN. In each stack of MASN, we generate multi-scale adaptive structure maps.

connection relationship to perform multi-person pose estimation. In this work, we adopt the method of establishing a model of the area between the two joints that have a connection relationship, and can completely learn the human skeleton structural information.

Multi-scale Structure Information. In the past two years, multi-scale network architecture has had a significant role in multiple tasks of vision[19] [20] [21]. In human pose estimation, the champion of COCO2017 [22] cascaded the RefineNet after FPN to better merge different scale feature information. Sun et al. [23] used multi-scale fully concolutional network to detect joints. Hourglass [5] has acquired information on multi-scale but has not conducted supervised learning at multi-scale. Ke et al. [8] added supervised learning of heatmaps at multi-scale of hourglass, but lacked more important body structural information. We supervise structural information at multi-scale of the network and describe human information in images more comprehensively.

3 Method

An overview of our framework is illustrated in Fig. 1. Our method achieves end-to-end learning which inputs RGB images and outputs joint heatmaps. We adopt the highly modularied Hourglass module[5] as the basic network structure to investigate the effect of Multi-scale Adaptive Structure Network(MASN). MASN add multi-scale adaptive structural information supervision on the basis of Hourglass module, and output adaptive heatmaps at the end. The ground-truth multi-scale adaptive structural information is represented by limb region,

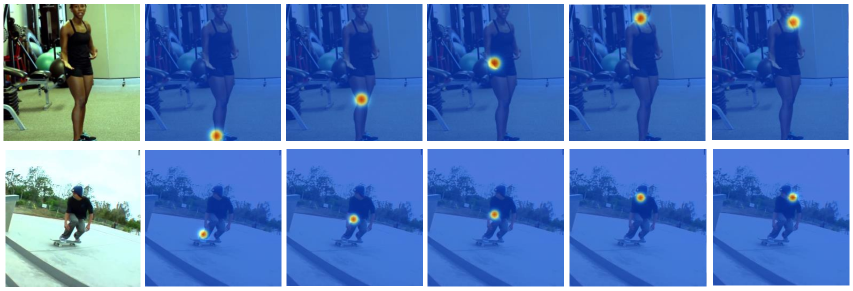


Fig. 2. Adaptive heatmaps. **Top**: large-scale human body uses large heatmaps. **Bottom**: small-scale human body uses small heatmaps.

which is the area between the two joints that have a connection relationship. The ground-truth adaptive heatmaps are generated using 2D Gaussian.

In this section, we specifically elaborate our approach in three areas: adaptive heatmaps, limb region(structural information), and multi-scale adaptive structure supervision.

3.1 Adaptive Heatmaps

In many pose estimation methods [2][7][22] including Hourglass, the ground-truth heatmap h_j^* of joint j is generated by 2D Gaussian:

$$h_j^*(p) = exp(-\frac{\|p - y_j\|^2}{2\sigma^2}) \tag{1}$$

where p is heatmap location, $j = 1, 2, ..., J$ is the index of each joint and J is the number of joints, y_j is the ground-truth at location p. σ controls the spread of the peak.

All existing methods use the same σ to generate heatmaps of the same kernel size, but obviously this is flawed. Although it is necessary to crop images according to the size of the human body area in the single-person pose estimation, the scale of the human body in the cropped image region is also inconsistent. In the far left side of Fig. 2, these are two training images of the same size($w \times h$), their human body scales are different. It is obviously not suitable to use the heatmaps of same kernel size to represent joints positions. We adpot adaptive heatmaps, which means that different σ is generated according to the size of the different body scale to control the kernel size of heatmaps. We chose the simplest and most appropriate method to generate adaptive heatmaps. Given an image($w \times h$), the maximum scale m of the human body shape is found by traversing according to the positions of whole body joints. Then the appropriate σ produced by linear interpolation,

$$\sigma = \frac{m}{max(w, h)}(\sigma_{max} - \sigma_{min}) + \sigma_{min} \tag{2}$$

where σ_{max} is the upper limit of σ, σ_{min} is the lower limit. In our experiments, $\sigma_{max} = 2.5$, $\sigma_{min} = 1.0$. Hence the ground-truth is $h^*_{j,\sigma}$. In Fig. 2, different kernel sizes of heatmaps are generated according to different body scale.

For each stack Hourglass i, the loss function L^i_h in [5] for training the model,

$$L^i_h = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{J}\sum_{j=1}^{J}\frac{1}{P}\sum_{p}\left\|\widehat{y}^i_j(p) - h^*_j(p)\right\|^2_2 \tag{3}$$

where N is the number of batch size, P denotes the number of all point p, $\widehat{y}^i_j$ denotes the predicted heatmap for joint j at stack Hourglass i. However, in our method, the large ground-truth heatmaps produce large loss function, the small ground-truth heatmaps produce small loss function. Therefore, it is necessary to normalize the loss function. For Gaussian functions,

$$\int exp(-\frac{x^2}{2\sigma^2}) = \sigma\sqrt{2\pi} \tag{4}$$

$$\sum(exp(-\frac{x^2}{2\sigma^2}))^2 = \sum exp(-\frac{x^2}{2(\frac{\sqrt{2}}{2}\sigma)^2}) \approx \int exp(-\frac{x^2}{2(\frac{\sqrt{2}}{2}\sigma)^2}) = \sigma\sqrt{\pi} \tag{5}$$

so the loss of every training sample $L^i_{h,n}$ is proportional to σ,

$$L^i_{h,n} \propto \sigma \tag{6}$$

In a batch, normalize each training sample. The final loss function is

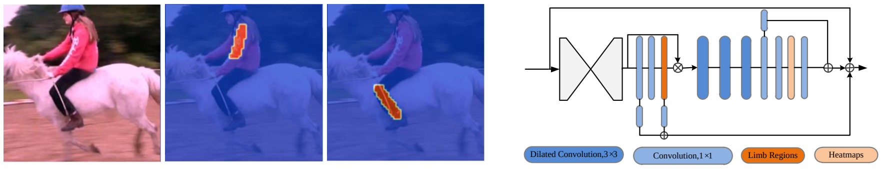
$$L^i_h = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{J\times\sigma_n}\sum_{j=1}^{J}\frac{1}{P}\sum_{p}\left\|\widehat{y}^i_j(p) - h^*_j(p,\sigma_n)\right\|^2_2 \tag{7}$$

where σ_n is generated by Equ. (2) for training sample n, $\widehat{y}^i_j$ denotes the predicted heatmap at stack i.

The adaptive heatmaps help to solve the problem that the heatmaps of joints detected is not accurate enough when the human body area is too large or the human body area is too small.

3.2 Limb Region

The human body structural information is the skeleton connection relationship of the human body. Obviously, the human body structure is one of the most important information for human pose estimation. Directly detecting joint heatmaps does not learn structural information primely, and how to characterize structural information is greatly important. Hence we propose a non-parametric representation called limb region to learn structural information, which denotes location information across the region of limb. In addition, our method can also be understood as an attention model, which is different from the previous method [17] focusing on the entire human body area. It breaks the attention of whole human



(a) (b)

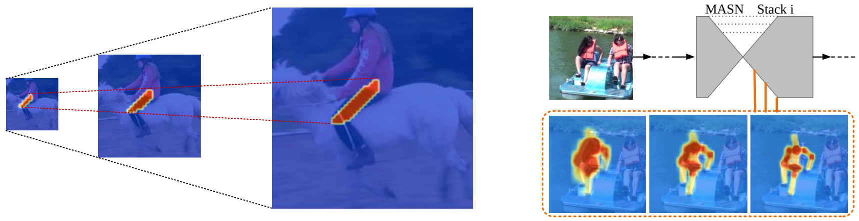
Fig. 3. **(a)**: Structural information is represented by limb region, e.g. left arm, left thigh. **(b)**: An illustration of the structural network. It consists of two parts: limb region detection and heatmaps detection.

body into the attention of limbs and makes it more accurate for the region of human body.

The limb region is a scalar region between the two joints that have a connection relationship. In the known human pose estimation datasets, the ground-truth of limb region is not given. It is necessary to split different limbs according to the ground-truth positions of joints. Consider a single limb l (as show in the middle/right of Fig. 3(a)), y_{j_a} and y_{j_b} are the ground-truth positions of joints adjacent to the limb l. The limb l is characterized by a rectangular box, the length is the distance $length$ between y_{j_a} and y_{j_b}, and the width is the set value $width$. If a point p lies on the rectangular box, the value of ground-truth limb region $r_l^*(p)$ is 1; for all other points, the value is 0. The ground-truth limb region at a map as,

$$r_l^*(p) = \begin{cases} 1, if\ p\ on\ limb\ l \\ 0, \quad otheriwise \end{cases} \tag{8}$$

We first used Hourglass for the detection of the limb region, followed by the joint heatmaps detection network, as shown in Fig. 3(b). Hourglass outputs 64×64 features, the channel is 256, after two 1×1 convolution to get the output of limb regions $L \times 64 \times 64$, where L is the number of limbs. Then concat Hourglass features with limb regions to get $(256+L)$-d features. Sequentially stacked three 3×3 dilated convolution[24][25] and two 1×1 convolution, joint heatmaps $J \times 64 \times 64$ are obtained. It is worth noting that after the first 3×3 dilated convolution, the feature dimension is reduced to 256. Similar to convolutional Pose Machine[4], the purpose of stacking three dilated convolution is to expand the receptive field so that the receptive field at the position of each joint can at least cover the neighboring limb region and make full use of the structural information of the limb region. However, convolutional Pose Machine uses large convolutional kernels$(9 \times 9, 11 \times 11)$ and generates a large amount of parameters. Significantly, we use the dilated convolution, which can both guarantee the receptive field and reduce the amount of parameters.



(a) (b)

Fig. 4. Multi-scale Adaptive Structure Supervision. **(a)**: Take the right thigh as an example to show the representation of the limb region on three scales. **(b)**: An example of Multi-scale limb regions output within a MASN. These limb regions are summed into a sigle map.

In the training phase of the limb region, we used the same loss function as heatmaps, the Mean Square Error loss. For each stack i, the loss function L_r^i is

$$L_r^i = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{L}\sum_{l=1}^{L}\frac{1}{P}\sum_{p}\left\|\widehat{r}_l^i(p) - r_l^*(p)\right\|_2^2 \tag{9}$$

where $\widehat{r}_l^i$ denotes the predicted limb region at stack i.

3.3 Multi-scale Adaptive Structure Supervision

In Sec. 3.2, we specifically state that human body structural information can be characterized by the limb region, and this section will expound human structural information at multi-scale. The first question, how do we generate multi-scale of the limb region? One possible way is downsampling the ground-truth limb region produced at the original scale. However, when the limb region is small, downsampling is extremely easy to distort, resulting in a discontinuous limb region. To address the limitation, we generate the ground-truth limb region on the corresponding scales by scaled the positions of joints. In our experiments, we generate three scales limb regions, as shown in Fig. 4(a). We enlarge the different scale of the limb region of the network to the original scale, which can be clearly found that the smaller scale, the more comprehensive the information described; the larger scale, more accurate, as shown in Fig. 4(b). For different scales of the limb region, the length is the distance between adjacent joints, but the width is inconsistent. In addition, adaptive width c is also needed for body scale, similar to adaptive σ in Equ. (2),

$$c_n^s = \frac{m_n}{max(w,h)}(c_{max}^s - c_{min}^s) + c_{min}^s \tag{10}$$

where c_n^s denotes the adaptive width of nth training sample at scale s, m_n denotes the maximum scale of body shape of nth training sample, c_{max}^s is the upper limit of limb region width at scale s, c^s_{min} is the lower limit of limb region width. In our experiments, $s = 1, \frac{1}{2}, \frac{1}{4}$, $c^1_{max} = 2.5$, $c^1_{min} = 1.0$, $c^{\frac{1}{2}}_{max} = 1.5$, $c^{\frac{1}{2}}_{min} = 0.75$, $c^{\frac{1}{4}}_{max} = 1.0$, $c^{\frac{1}{4}}_{min} = 0.5$.

Similar to the adaptive sigma, the loss function of the adaptive limb region needs to be normalized, and the loss functions of different scales are also normalized. The multi-scale adaptive limb region loss function as,

$$L_r^i = \sum_s \frac{s^2}{N} \sum_{n=1}^{N} \frac{1}{L \times c_n^s} \sum_{l=1}^{L} \frac{1}{P^s} \sum_{p^s} \left\| \widehat{r}_l^i(p^s) - r_l^*(p^s, c_n^s) \right\|_2^2 \tag{11}$$

where $\widehat{r}_l^i(p^s)$ denotes the predicted limb region position p^s of scale s, $r_l^*(p^s, c_n^s)$ denotes the ground-truth limb region which generates according to adaptive width c_n^s, P^s denotes the number of all point at scale s. It should be noted that s^2 is a value set to balance the loss of limb region between multi-scale. In addition to the learning of the limb region, each complete MASN also needs to detect the joint heatmaps, so the complete loss function is

$$L = \sum_i L^i = \sum_i \alpha L_h^i + \beta L_r^i \tag{12}$$

where α and β is to balance the two loss functions. In our experiments, $\alpha = 1$, $\beta = 0.1$.

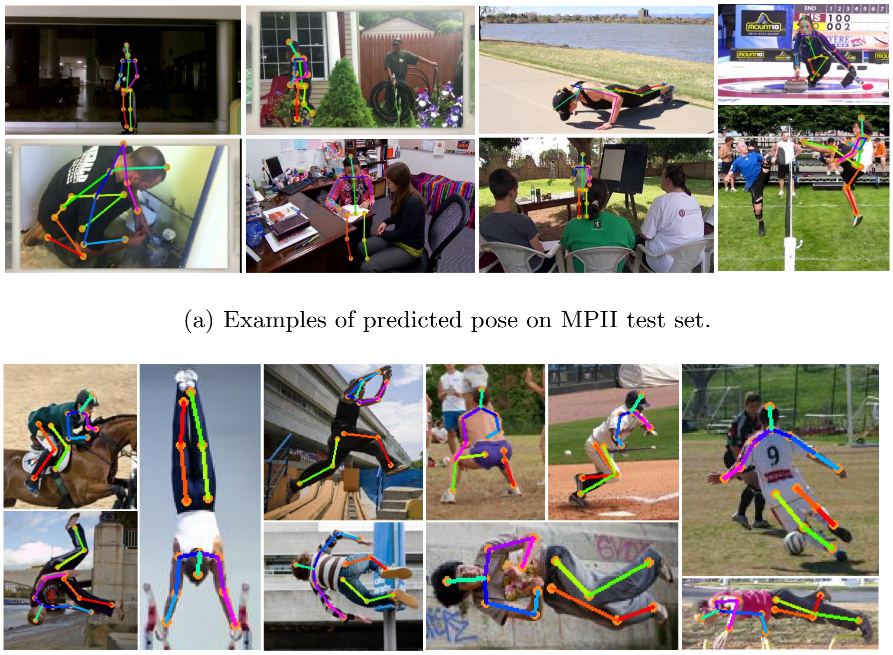
4 Experiments

4.1 Implementation Details

Dataset. We evaluate our method on two widely benchmark datasets, MPII Human Pose [9] and Leeds Sports Poses (LSP) [26] and its extended training dataset. MPII Human Pose dataset includes around 25K images containing over 40K people with annotated body joints, which covers a wide range of human activities. LSP dataset is composed of 12K images with challenging poses in sports activities.

Data Augmentation. We crop the images with the annotated body center and scale, and resize to $256 * 256$. Then we augment the images by performing random scaling factor between 0.75 and 1.25, horizontal flipping, and rotating across $\pm 30^\circ$. In addition, we add color noise to make the model more robust. During testing, we crop the images with approximate location and scale of each person for MPII dataset. Since the LSP dataset has no location and scale, we use the image center and image size. All our experimental results is conducted on 6-scale image pyramids with flipping.

Experiment Settings. Due to the shortage of hardware devices, we only stack 4 MASNs as our complete network structure. Our method is implemented using PyTorch open-source framework and we use RMSprop [27] algorithm to optimize the network on 1 NVIDIA GTX 1080Ti GPU with a mini-batch size of 6 for 200 epochs. The initial learning rate is 1.5×10^{-4} and is dropped by 10 at 110th and the 160th epoch.



(a) Examples of predicted pose on MPII test set.

(b) Examples of predicted pose on LSP test set.

Fig. 5. Qualitative results on two datasets.

4.2 Results

Evaluation measure. We use the Percentage Correct Keypoints (PCK) [33] measure on the LSP dataset, which reports the percentage of predicted joint position within a normalized distance of the ground-truth. For MPII dataset, the PCKh [9], as an improved PCK, uses the head size for normalization, making the evaluation more accurate.

MPII dataset. Our results are reported in Table 1 using PCKh@0.5 measure, 50% of the head size for normalization. Our model is trained by dividing the MPII complete training data set into a training set and a validation set as [28]. We reach a test score of 91.7%, which is 0.8% higher than 8-stack Hourglass [5]. Obviously, it is a 0.2% improvement over the improved attention mechanic [17] based on Hourglass, which is also a stack of 8 modules. The results of Yang et al.[6] are basically the same as ours, but comparing the model parameters, our model has fewer parameters. In particular, for the most challenging joints, our method achieves 1.0% and 1.0% improvements on wrist and ankle compared with 8-stack Hourglass. Examples of quantitative results are shown in Fig. 5(a).

LSP dataset. We add MPII training data to the LSP dataset and its extended training dataset in a similar way as before [4][17]. Table 2 summarizes the PCK scores at threshold of 0.2 with PC (Person-Centric) annotations. Our

Table 1. Comparison MPII dataset results(PCKh@0.5 score) and model parameters

Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean	Para.(M)
Carreira et al.[3]	95.7	91.7	81.7	72.4	82.8	73.2	66.4	81.3	–
Tompson et al.[28]	96.1	91.9	83.9	77.8	80.9	72.3	64.8	82.0	–
Gkioxary et al.[29]	96.2	93.1	86.7	82.1	85.2	81.4	74.1	86.1	–
Wei et al.[4]	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5	–
Bulat et al.[30]	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7	–
Newell et al.[5]	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9	23.7
Chu et al.[17]	98.5	96.3	91.9	88.1	90.6	88.0	85.0	91.5	>23.7
Yang et al.[6]	98.4	96.5	91.9	**88.2**	**91.1**	**88.6**	**85.3**	**91.8**	26.9
Ours	**98.5**	**96.7**	**92.1**	88.1	90.9	88.2	84.6	91.7	**21.6**

Table 2. Comparison LSP dataset results(PCK@0.2 score) and model parameters

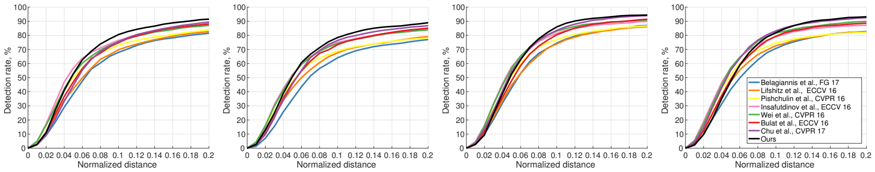
Method	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean	Para.(M)
Belagiannis et al.[31]	95.2	89.0	81.5	77.0	83.7	87.0	82.8	85.2	–
Lifshitz et al.[32]	96.8	89.0	82.7	79.1	90.9	86.0	82.5	86.7	–
Wei et al.[4]	97.8	92.5	87.0	83.9	91.5	90.8	89.9	90.5	–
Bulat et al.[30]	97.2	92.1	88.1	85.2	92.2	91.4	88.7	90.7	–
Chu et al.[17]	98.1	93.7	89.3	86.9	93.4	94.0	92.5	92.6	>23.7
Ours	**98.1**	**94.6**	**91.5**	**89.0**	**94.2**	**94.4**	**93.2**	**93.6**	**21.6**

method increased by 1.0% compared to the previous method, and in particular, there was a 2.2% and 2.1% increase in elbow and wrist. In Fig. 6, we show that our method performs significantly better than before when the normalized distance is greater than 0.08. Examples of predicted pose are demonstrated in Fig. 5(b).

4.3 Ablation Study

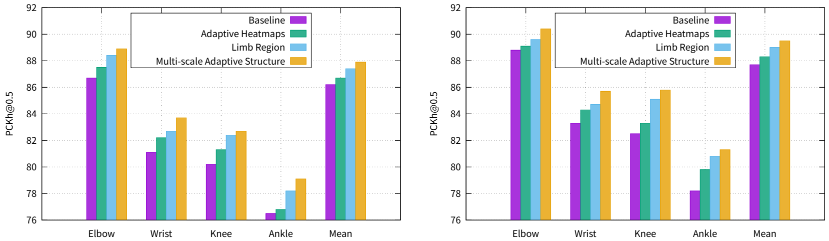
In order to study whether the method mentioned in Sec. 3 is really effective for pose estimation. We conducted a series of experiments on the MPII validation set [28]. First, we use 1-stack Hourglass as our baseline for comparison. We need to analyze each part of our approach, including: Adaptive Heatmaps, the Limb Region (structural information), and Multi-scale Adaptive Structure (1-stack MASN). We use the most challenging joints and the overall score as the basis for evaluation to compare with 1-stack Hourglass score of 86.2% at PCKh@0.5, as shown in Fig. 7(a).

Adaptive Heatmaps. In order to explore whether the size of the heatmap affect the model effect, we first evaluate the adaptive heatmaps method. By comparing with baseline, we got a PCKh@0.5 score of 86.7% with a 0.5% improvement. The most significant increase in knees is 1.1%.



(a) Elbow (b) Wrist (c) Knee (d) Ankle

Fig. 6. PCK curves on the LSP dataset on some challenging body joints.



(a) 1-stack (b) 2-stack

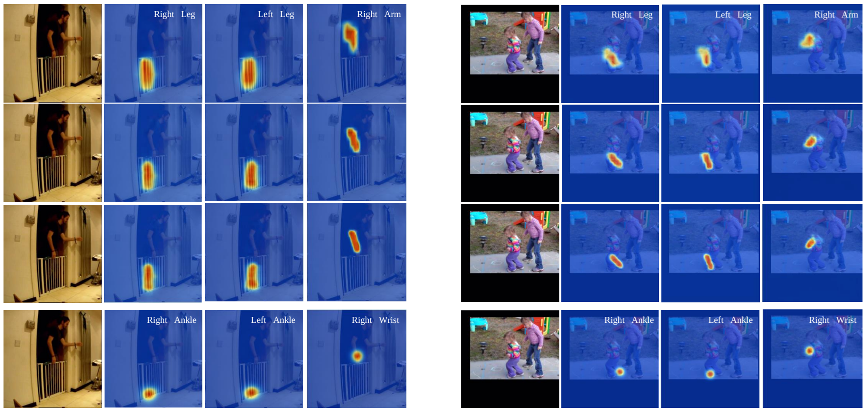
Fig. 7. Ablation Study. Comparison PCKh@0.5 score on MPII validation set. **(a)**: 1-stack Hourglass as baseline. **(b)**: 2-stack Hourglass as baseline.

Limb Region. The Limb Region, as an expression of the human body structure, performs very well in our experiments. Compared to baseline, this method achieves a PCKh@0.5 score of 87.4%, with significant effects at several challenging joints.

Multi-scale Adaptive Structure. We supervise the limb region at $\frac{1}{4}$, $\frac{1}{2}$, and original scale. Due to the inconsistency of the human body size in the image, we adopt an adaptive approach for the limb region at each scale and adaptive heatmaps for joints. At the same time, according to the method in Sec. 3, the loss function is normalized. In the end, we get a 87.9% PCKh@0.5 score.

1-stack Hourglass vs. 1-stack MASN. We obtain a PCKh@0.5 score of 87.9% by stacking one MASN, which is a 1.7% increase over 1-stack Hourglass. We visualize some of the results and found that it has a significant effect on the illumination, occlusion and other complex situations.

The comparison of 2-stack. Similarly, we use 2-stack Hourglass as our baseline for comparison. Analyzing Adaptive Heatmaps, the Limb Region (structural information), and Multi-scale Adaptive Structure (1-stack MASN), as shown in Fig. 7(b). Same as 1-stack, each part of our approach has a better score than baseline, and 2-stack MASN achieves a PCKh@0.5 score of 89.5%, a 1.8% increase over 2-stack Hourglass.



(a) (b)

Fig. 8. Two examples of qualitative analysis. 1st row to 3rd row: $\frac{1}{4}$, $\frac{1}{2}$, and original scale of adaptive limb region. 4th row: adaptive heatmaps of joints. **(a)**: Large body size, illumination and occlusion. **(b)**: Small body size, occlusion.

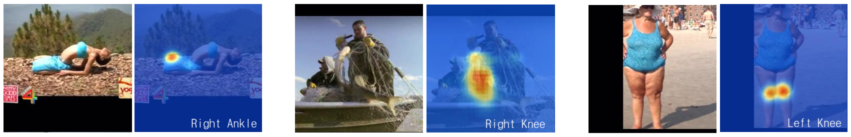
4.4 Discussion

Some of the visualized examples, as shown in Fig. 5, illustrate that our method has significant results for occlusion, illumination, crowding, complex backgrounds, rare pose, and so on. In this section, we first conduct a specific analysis about the performance of MASN.

Qualitative Analysis. The MASN module detects the adaptive limb region at $\frac{1}{4}$, $\frac{1}{2}$, and the original scale, and then detects the adaptive heatmaps of joints, as shown in Fig. 8. The limb region in small-scale can provide context information for the limb region refinement in large-scale. The limb region fully learns the structure information of the human body, which plays a key role in the prediction of invisible joints under the conditions of occlusion, illumination and so on. Moreover, the self-adaptive method of the limb regions and the heatmaps proposed by us is also crucial for different sizes of human body joints. It can help to predict appropriate results for human bodies of different sizes, thereby improving accuracy.

Model Parameters. In Sec. 4.2 and 4.3, we specifically illustrate the performance advantages of our approach over Hourglass, and we also compare the number of parameters. The number of parameters for stacking 4 MASN modules is 21.6M, and the number of parameters for stacking 8 Hourglass modules is 23.7M, as shown in Table 1. We improve performance while reducing the amount of parameters.

PoseTrack dataset [34]. Our approach may fail under some extreme conditions, as shown in Fig. 9. For example, too abnormal human pose, heavy occlusion, and abnormal human body. Further analysis, we find that there are relatively few such data in the dataset, and our model has not yet perfected this



(a) (b) (c)

Fig. 9. Failure caused by (a) occlusion and rare pose, (b) overlapping and occlusion, (c) abnormal body(overweight).

kind of situation. So we use a larger dataset, PoseTrack dataset [34], to train ours model. Since the PoseTrack dataset does not have the label ”person scale” like MPII dataset, we get the ”person scale” directly from the marked joint points. Due to the limitations of the device, our model is not pre-trained on the COCO dataset [35] like other methods [36], and does not take advantage of the relationship between frames, but directly train on the labeled images. Although we don’t get result as good as [36], the focus of our future work will be here, that is, combines our methods with human detection, end-to-end multi-person pose estimation, and the ability to use the continuous information of the video to estimate a smooth human body.

5 Conclusions

In this paper, we discuss the importance of variant sizes of the joint heatmaps for human pose estimation. We propose a novel method to address this issue by adapting the kernel size of joint heatmaps to the human scale of the input images in the training stage. A normalization strategy of how to perform the adaption is deduced. Besides, we introduce a novel limb region representation to learn the human pose structural information. Both the adaptive joint heatmaps as well as the limb region representation are combined together to construct a novel neural network, which is named **Multi-scale Adaptive Structure Network (MASN)**. We demonstrate that our method could obtain the state-of-the-art human pose estimation results and outperform the most existing methods.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (No. 61806054, 6170320, 61728103 and 61671151), Jiangsu Province Science Foundation for Youths (No. BK20180355 and BK20170812) and Foundation of Southeast University (No. 3208008410 and 1108007121).

References

1. Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2014) 1653–1660
2. Pfister, T., Charles, J., Zisserman, A.: Flowing convnets for human pose estimation in videos. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 1913–1921
3. Carreira, J., Agrawal, P., Fragkiadaki, K., Malik, J.: Human pose estimation with iterative error feedback. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 4733–4742
4. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 4724–4732
5. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision, Springer (2016) 483–499
6. Yang, W., Li, S., Ouyang, W., Li, H., Wang, X.: Learning feature pyramids for human pose estimation. In: The IEEE International Conference on Computer Vision (ICCV). Volume 2. (2017)
7. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR. Volume 1. (2017) 7
8. Ke, L., Chang, M.C., Qi, H., Lyu, S.: Multi-scale structure-aware network for human pose estimation. arXiv preprint arXiv:1803.09894 (2018)
9. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on computer Vision and Pattern Recognition. (2014) 3686–3693
10. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. International journal of computer vision **61** (2005) 55–79
11. Ramanan, D.: Learning to parse images of articulated bodies. In: Advances in neural information processing systems. (2007) 1129–1136
12. Buehler, P., Everingham, M., Huttenlocher, D.P., Zisserman, A.: Long term arm and hand tracking for continuous sign language tv broadcasts. In: Proceedings of the 19th British Machine Vision Conference, BMVA Press (2008) 1105–1114
13. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE (2009) 1014–1021
14. Pishchulin, L., Andriluka, M., Gehler, P., Schiele, B.: Strong appearance and expressive spatial models for human pose estimation. In: Computer Vision (ICCV), 2013 IEEE International Conference on, IEEE (2013) 3487–3494
15. Taylor, G.W., Fergus, R., Williams, G., Spiro, I., Bregler, C.: Pose-sensitive embedding by nonlinear nca regression. In: Advances in Neural Information Processing Systems. (2010) 2280–2288
16. Chen, X., Yuille, A.L.: Articulated pose estimation by a graphical model with image dependent pairwise relations. In: Advances in neural information processing systems. (2014) 1736–1744
17. Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A.L., Wang, X.: Multi-context attention for human pose estimation. arXiv preprint arXiv:1702.07432 **1** (2017)
18. Chu, X., Ouyang, W., Li, H., Wang, X.: Structured feature learning for pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 4715–4723

19. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR. Volume 1. (2017) 4
20. Gidaris, S., Komodakis, N.: Object detection via a multi-region and semantic segmentation-aware cnn model. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 1134–1142
21. Girshick, R., Iandola, F., Darrell, T., Malik, J.: Deformable part models are convolutional neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. (2015) 437–446
22. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. arXiv preprint arXiv:1711.07319 (2017)
23. Sun, K., Lan, C., Xing, J., Zeng, W., Liu, D., Wang, J.: Human pose estimation using global and local normalization. arXiv preprint arXiv:1709.07220 (2017)
24. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
25. Yang, Z., Hu, Z., Salakhutdinov, R., Berg-Kirkpatrick, T.: Improved variational autoencoders for text modeling using dilated convolutions. arXiv preprint arXiv:1702.08139 (2017)
26. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. (2010)
27. Tieleman, T., Hinton, G.: Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural networks for machine learning **4** (2012) 26–31
28. Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C.: Efficient object localization using convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 648–656
29. Gkioxari, G., Toshev, A., Jaitly, N.: Chained predictions using convolutional neural networks. In: European Conference on Computer Vision, Springer (2016) 728–743
30. Bulat, A., Tzimiropoulos, G.: Human pose estimation via convolutional part heatmap regression. In: European Conference on Computer Vision, Springer (2016) 717–732
31. Belagiannis, V., Zisserman, A.: Recurrent human pose estimation. In: Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on, IEEE (2017) 468–475
32. Lifshitz, I., Fetaya, E., Ullman, S.: Human pose estimation using deep consensus voting. In: European Conference on Computer Vision, Springer (2016) 246–260
33. Yang, Y., Ramanan, D.: Articulated human detection with flexible mixtures of parts. IEEE transactions on pattern analysis and machine intelligence **35** (2013) 2878–2890
34. Andriluka, M., Iqbal, U., Milan, A., Insafutdinov, E., Pishchulin, L., Gall, J., Schiele, B.: Posetrack: A benchmark for human pose estimation and tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 5167–5176
35. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision, Springer (2014) 740–755
36. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. arXiv preprint arXiv:1804.06208 (2018)