Pattern Recognition Letters 154 (2022) 1–6

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Scale-aware heatmap representation for human pose estimation

Han Yu, Congju Du, Li Yu*

School of Electronic Information and Communication, Huazhong University of Science and Technology, Wuhan 430000, China

ARTICLE INFO

Article history:
Received 25 May 2021
Revised 14 December 2021
Accepted 27 December 2021
Available online 29 December 2021

Edited by Jiwen Lu

MSC:
41A05
41A10
65D05
65D17

Keywords:
Human pose estimation
Multi-person
Heatmap representation

ABSTRACT

The performance of multi-person pose estimation is seriously affected by scale variation. Extensive works have been devoted to reducing the effect by modifying convolutional network structure or loss function, but little attention has been paid to the problem in the construction of heatmaps. In this paper, we focus on the scale variation of keypoints within heatmap generation and propose a novel method called scale-aware heatmap generator, which constructs a customized heatmap for each type of keypoints based on their relative scales. In addition, we design a weight-redistributed loss function to facilitate the detection of keypoints that are hard to identify. Our approach outperforms the baseline by nearly 2.5% in average precision and performs on par with the state-of-the-art result in bottom-up pose estimation with multi-scale testing (69.4% AP) on the COCO test-dev dataset.

1. Introduction

Human pose estimation, which aims to detect and locate all human anatomical keypoints from given images or video sequences, is one of the core tasks in computer vision. It has captured a lot of research interests due to its significance in various applications, such as human action recognition and pedestrian tracking. Traditional methods [1,2] are mostly based on graphical models or pictorial structures, but they are vulnerable and frangible in complicated situations (e.g., occlusion). Thanks to the rapid development of deep learning, the employment of convolutional neural networks [3,4] has made substantial progress in this field.

Recently, the approaches for multi-person pose estimation can be divided into top-down and bottom-up methods. The top-down method first obtains person instances based on object detectors and then performs single person pose estimation individually. The bottom-up method first detects all keypoints in the image and then groups them into the corresponding people. The top-down method usually achieves higher accuracy as it crops and resizes the sub-image which contains a single person to a fixed scale for pose estimation. However, its performance tightly depends on the quality of the object detectors and its runtime is almost linearly related to the number of people in the image. In contrast, the bottom-up method can achieve real-time inference regardless of the number of people and performs better in more complicated situations (e.g., crowd scenarios).

Heatmap representation is widely used to encode the location of keypoints. An illustration of a standard heatmap is shown in Fig. 1(b). The heatmap can be defined as a confidence map generated from the Gaussian kernel centered at each annotated keypoint, and the heat value indicates the probability whether there exists the target keypoint. Correspondingly, the keypoint positions can be obtained by identifying the local maximums in the heatmap. Compared with prior approaches [3], heatmap contains more spatial information near the keypoints, boosting the performance of convolutional neural networks. Therefore, most of the mainstream methods (e.g., HigherHRNet [5], RSN [6], DarkPose [7]) use heatmap as the default configuration. However, little attention has been paid to the scale variation problem within the generation of heatmaps. The standard heatmap adopts a fixed Gaussian kernel for all types of keypoints, which may cause confusion as different types of keypoints have distinct scales. As shown in Fig. 1(b), two kinds of keypoints with different sizes (like eye and hip) are consistent in heatmap, but their influence on adjacent keypoints is completely different. It is unreasonable to use the same Gaussian kernel for each type of keypoints to generate heatmaps.

Motivated by the above observation, we propose a novel algorithm, namely scale-aware heatmap generator, to tackle the

* Corresponding author.
E-mail address: hustlyu@hust.edu.cn (L. Yu).

https://doi.org/10.1016/j.patrec.2021.12.018
0167-8655/

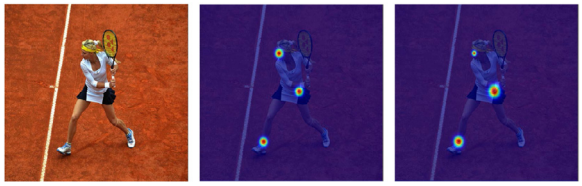

Fig. 1. Illustration of heatmaps encoded with Gaussian kernels. (a) Original image, (b) standard heatmap, and (c) scale-aware heatmap with left eye, left hip and right ankle. Compared with the standard heatmap, the proposed scale-aware heatmap is more informative as it contains the scale information of keypoints.

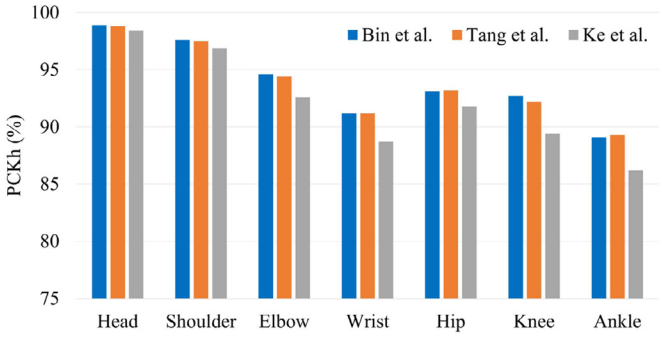

Fig. 2. Human pose estimation evaluation on MPII dataset [12]. PCKh metric [12] is used to measure accuracy of the localization. We compare the evaluations in different types of keypoints on three state-of-the-art methods [13–15]. It can be found that keypoints with small scales (e.g., wrist and ankle) are with lower PCKh values, indicating that they are harder to be detected.

scale variation of keypoints within heatmap generation. We have counted the relative scale proportions of all types of keypoints according to [8], and calculated adaptive variance of Gaussian kernel based on the relative scales. Specifically, we form a unified variance and compute the suitable variance for each type of keypoints proportionally, thus we can incorporate the scale information into heatmaps.

In addition, plenty of experimental results show that difference in the scale of keypoints may also affect the localization accuracy (shown in Fig. 2). There exists a detection imbalance problem that keypoints with smaller scales are harder to detect. To deal with this problem, the design idea of focal loss [9] is adopted in the network. We build a weight-redistributed loss and it can automatically adjust the contribution between disparate keypoints during the training process.

We use HRNet [10] as backbone and follow [11] to build a bottom-up pose estimation system. We evaluate our method on COCO val2017 dataset without multi-scale testing and it achieves 67.5% AP, outperforming baseline by 2.5% AP. Our model achieves 69.4% AP on COCO test-dev dataset, which is comparable with those state-of-the-art methods. We have performed several ablation experiments to show the effectiveness of each component in our approach.

The main contributions can be summarized as follows:

- We propose a scale-aware heatmap generator (SAHG) to tackle the scale variation of keypoints within the construction of heatmaps. Different from the standard heatmap that adopts a unified kernel for all types of keypoints, our SAHG will generate a customized heatmap for each type of keypoints according to their relative scales;
- To alleviate the detection imbalance problem and facilitate the detection of hard keypoints, we design a weight-redistributed loss function, which regulates the contribution between different keypoints automatically and thus improves the estimation accuracy.

2. Related work

Extensive efforts have been made to handle the scale variation of human bodies within pose estimation, which can be classified into two categories. One is to modify the structure of convolutional neural network referring to the image pyramid structure, another is to incorporate the impact of scale in the construction of loss functions.

Feature pyramid. FPN [16] exploits feature pyramids to extract multi-scale features, designing a pyramidal hierarchy network followed by a number of works [17,18]. Hourglass [19] employs a symmetrical network structure to combine different scale representations, which is regarded as a milestone in pose estimation. CPN [17] designs two sub-networks: GlobalNet and RefineNet. The GlobalNet adopts pyramid structure to extract the spatial and semantic information of the image and the RefineNet is designed to address those hard-detected keypoints. MSPN [20] comes up with a cross stage aggregation strategy to reduce information loss and a coarse-to-fine supervision strategy to improve localization accuracy. Simple Baseline [21] notices the significance of high-resolution representations for exact coordinates, utilizing deconvolution modules to generate high-resolution features. Inspired by [21], HRNet [10] designs a novel structure to maintain high-resolution representations over the network, resulting in a high performance with the fusion of various scale features. HigherHRNet [5] further takes advantage of the deconvolution module to generate higher resolution representations based on [10], achieving state-of-the-art in bottom-up pose estimation.

Loss function. PifPaf [22] indicates that localization errors of keypoints may be affected by different scales of people in various ways, so it constructs a novel loss function and injects a scale dependence into it. CrowdPose [23] incorporates an attenuation factor into the heatmap loss function, which aims to strengthen the target joint response and suppress the interference joint response. SimplePose [24] observes that plenty of easy-detected samples may attract most of the attention in training, thus impeding the network from learning hard samples. To enhance the capacity of the network, it follows [9] to build a focal L2 loss function to tackle the imbalance problem.

Different from the two categories mentioned above, our work concentrates on the scale variation of keypoints within heatmap generation. We find that it is unreasonable to adopt the same Gaussian kernel to generate heatmaps as different kinds of keypoints have distinct scales. SWAHR [25] tries to handle this problem by learning scale maps using convolutional neural networks. Unlike it, the proposed SAHG takes full advantage of the statistical prior knowledge and constructs a rule to generate heatmaps in accordance with the scales of keypoints.

3. Proposed methods

According to [5,10,22,24], heatmaps are widely leveraged as coordinate representations to learn a robust model. Since heatmaps can provide rich spatial information during training and reflect the keypoint location probability for detection, reasonable heatmaps have a great effect on the performance of human pose estimation.

However, how to generate more suitable heatmaps has rarely been considered systematically by existing works. To address this problem, we propose SAHG and the whole structure is demonstrated in Fig. 3. The SAHG could incorporate scale information into heatmaps and generate more precise heatmaps for each type of keypoints based on their scales.

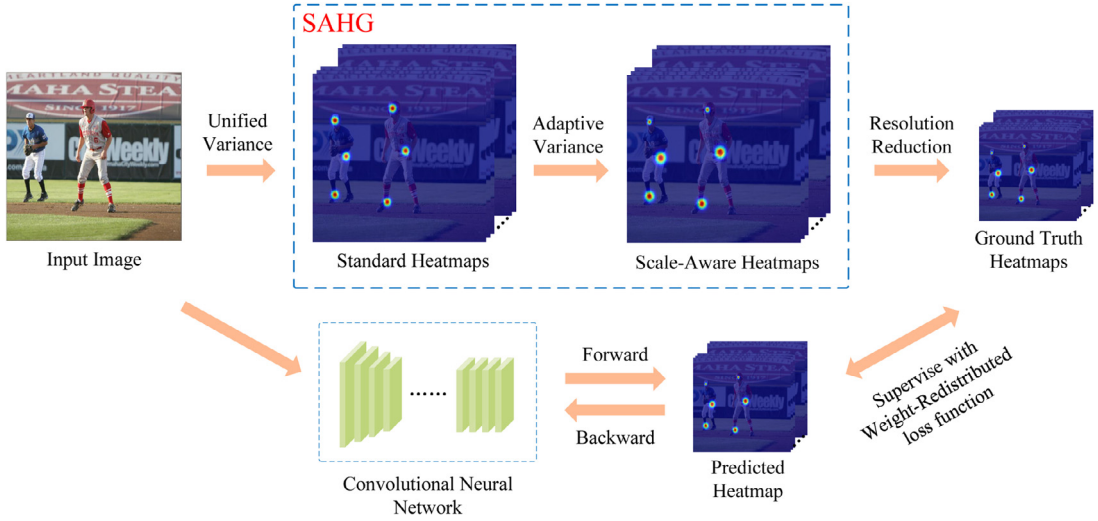

Fig. 3. An overview of the proposed SAHG embedded in human pose estimation. The top row shows the generation of ground truth heatmaps. We first set up a unified variance and then calculate adaptive variances for different types of keypoints based on their relative scales. The bottom row shows the training process supervised with weight-redistributed loss function.

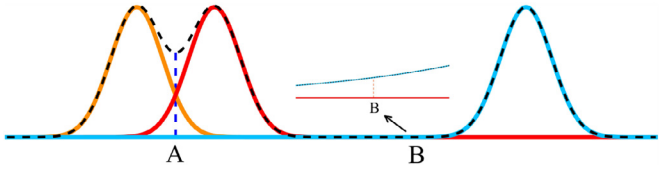

Fig. 4. Illustration of heatmap aggregation in multi-person pose estimation. The orange, red, and blue solid lines represent the same type of keypoint encoded by Gaussian kernels respectively. The black dotted line corresponds to their aggregation. Position A shows that multiple neighboring Gaussian distributions will interfere with each other, making them hard to identify. Position B shows that pixels far away from all peaks still receive an unignorable value. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

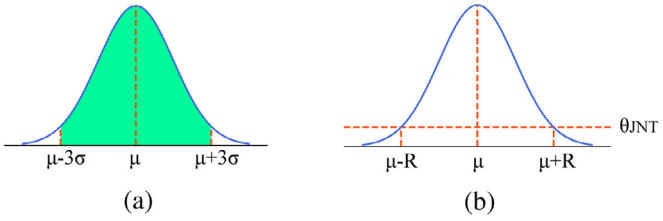

Fig. 5. The schematic diagram for calculating truncated radius. (a) Improved standard heatmap generator guarantees that the probability of points within the truncated radius is 99.7% based on PauTa criterion. (b) SAHG guarantees that the heat values of points within the truncated radius all above θ_{JNT}.

3.1. Improved standard heatmap generator

For two-dimensional human pose estimation, the standard heatmap generator maps each type of keypoint coordinates to a common Gaussian distribution heatmap, which can be represented as:

$$\mathcal{H}^i_{st}(\boldsymbol{x};\boldsymbol{\mu},\Sigma)=\sum_{j=1}^{N}\lambda\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \tag{1}$$

where $\lambda=1/(2\pi|\Sigma|^{\frac{1}{2}})$, N represents the number of human instance, i represents the type of keypoints, $\boldsymbol{x}$ denotes the pixel location and $\boldsymbol{\mu}$ represents the Gaussian mean vector.

The covariance matrix Σ is set as a diagonal matrix and σ represents the standard deviation of the Gaussian function:

$$\Sigma=\begin{bmatrix}\sigma^2 & 0\\ 0 & \sigma^2\end{bmatrix}. \tag{2}$$

Standard heatmap generator is widely used in single-person pose estimation, but its performance is seriously degraded when applied to multi-person pose estimation. Due to the unknown number of people, each pixel on heatmap may be affected by multiple Gaussian distributions. We can notice that the pixels far away from the ground truth keypoint coordinates still receive a certain value, which could adversely affect the precise results (illustrated in Fig. 4).

To deal with the above problem, an improved standard heatmap generator came up to restrict the spread of each Gaussian distribution. It could calculate a reasonable confidence interval to truncate the upper and lower bounds of Gaussian distribution, and the PauTa criterion is utilized here to obtain the truncated radius of the Gaussian distribution. The improved heatmap generator can be expressed as:

$$\mathcal{H}^i_{ist}(\boldsymbol{x};\boldsymbol{\mu},R)=\begin{cases}\mathcal{H}^i_{st}(\boldsymbol{x};\boldsymbol{\mu},\Sigma), & d_C(\boldsymbol{x};\boldsymbol{\mu})\le R\\ 0, & \text{else},\end{cases} \tag{3}$$

where $d_C(\cdot)$ is the Chebyshev distance defined as the absolute magnitude of the maximum differences between coordinates, R is the truncated radius that limits the spread of the Gaussian distribution. Here, R is set to 3σ to ensure that the probability of points within the truncated radius is sufficiently high (99.7%) based on PauTa criterion. The schematic diagram is shown in Fig. 5(a).

3.2. Scale-aware heatmap generator

We rethink the standard heatmap generator and discover that it simply adopts a unified Gaussian kernel for all keypoints to generate heatmaps, ignoring the inherent-scale feature of keypoints within the human body. It can be easily noticed that the size of hip keypoint is often several times larger than that of the eye, which may cause confusion as we use the same kernel to encode their locations.

Hence, it is reasonable to assign different types of keypoints with different Gaussian kernels, and those kernels can be determined by adjusting their variance values. For instance, the variance adjusted to fit hip keypoint is not suitable for eye keypoint, while

the standard heatmap generator does not fully consider this situation. Accordingly, we consider that it is proper to allocate the variance in accordance with the scale of keypoints to generate heatmaps.

To obtain the adaptive variance of different types of keypoints, we have performed statistics on the relative scales of disparate keypoints as reported by [8]. Then we design a basic variance value and multiply it by the relative proportion of scales to calculate the suitable variance value for each type of keypoints. We set a limit on the minimum variance value to avoid the value being too small to encode the spatial information near the keypoint. In addition, we propose a threshold named just noticeable threshold (JNT) to regulate the spread of Gaussian distribution dynamically. The threshold can be regarded as distinction between foreground and background pixels, and the heat value belows it will be set to zero directly as shown in Fig. 5(b). Therefore, as we mentioned in Section 3.1, the truncated radius of the Gaussian distribution can be calculated adaptively by fixing the JNT without additional parameters.

We can construct an equation for the association of three key variables, which are the truncated radius R, the standard deviation of the Gaussian kernel σ, and the JNT θ_{JNT}. For simplicity, we use the analysis of one-dimensional Gaussian distribution as an illustrative example:

$$\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{R^2}{2\sigma^2}\right)=\theta_{JNT}, \tag{4}$$

when the standard deviation σ and the JNT θ_{JNT} are determined, the truncated radius R will be calculated by (4):

$$R=\sqrt{2\sigma^2\ln((2\pi)^{1/2}\sigma\theta_{JNT})^{-1}}. \tag{5}$$

In this way, the influence region of Gaussian distribution can be controlled with the variance, thus we can generate more precise heatmaps by allocating the variance in accordance with the scale of keypoints. The SAHG can be expressed as follows:

$$\mathcal{H}^i_{sa}(\boldsymbol{x};\boldsymbol{\mu},\Sigma_i)=\begin{cases}\mathcal{H}^i_{st}(\boldsymbol{x};\boldsymbol{\mu},\Sigma_i), & \mathcal{H}^i_{st}(\boldsymbol{x})\geq\theta_{JNT}\\ 0, & \text{else},\end{cases} \tag{6}$$

$$\Sigma_i=\begin{bmatrix}\sigma_i^2 & 0\\ 0 & \sigma_i^2\end{bmatrix}. \tag{7}$$

The Σ_i is adaptive to each type of keypoints and the corresponding σ_i^2 can be obtained by the following equation:

$$\sigma_i^2=max(s_i,s_{thr})*\sigma^2/s_{thr} \tag{8}$$

where s_i is the relative scale of ith keypoint, s_{thr} is the scale threshold to limit the minimum scale and σ^2 is the basic variace value adopted by standard heatmap generator.

3.3. Weight-redistributed loss function

MSE loss function is frequently used to measure the similarity between the ground truth and the estimated heatmaps. However, through our analysis of the experimental results (shown in Fig. 2), the localization accuracy can be easily affected by the difference of keypoints which has not been well considered in MSE loss function. To solve this problem, we build a weight-redistributed loss, which can automatically reallocate the contribution between different keypoints in loss function.

The points with high heat values in heatmaps are supposed to be easily classified by the network while those with value close to θ_{JNT} are hard to distinguish well. As plenty of easy-classified samples may impede the network from learning hard samples [24], the proposed weight-redistributed loss assigns higher weights to those hard samples and reduces the contributions of easy samples. Thus we can supervise the network to learn better to classify those hard points during the training process. In practice, weight-redistributed loss generates weight maps $\hat{\mathcal{H}}_f$ to assign weights to each pixel in heatmaps based on their heat values, which can be expressed as:

$$\hat{\mathcal{H}}^i_f(\boldsymbol{x})=\begin{cases}e^{\hat{\mathcal{H}}^i(\boldsymbol{x})+\alpha_1}+\beta_1, & \hat{\mathcal{H}}^i(\boldsymbol{x})\leq\theta_{JNT}\\ -e^{\hat{\mathcal{H}}^i(\boldsymbol{x})+\alpha_2}+\beta_2, & \text{else},\end{cases} \tag{9}$$

where α_1,α_2 and β_1,β_2 are compensation factors, $\boldsymbol{x}$ denotes the pixel location, i represents the type of keypoints, $\hat{\mathcal{H}}^i(\boldsymbol{x})$ means the value located at $\boldsymbol{x}$ in the ith estimated heatmap $\hat{\mathcal{H}}^i$, θ_{JNT} is set to distinguish the foreground and background pixels as mentioned in Section 3.2.

With the weight maps $\hat{\mathcal{H}}_f$ and element-wise multiplication $\circ$, the loss between the ground truth heatmaps $\mathcal{H}_{sa}$ and the estimated heatmaps $\hat{\mathcal{H}}$ can be calculated as:

$$Loss=mean((\hat{\mathcal{H}}-\mathcal{H}_{sa})\circ(\hat{\mathcal{H}}-\mathcal{H}_{sa})\circ\hat{\mathcal{H}}_f). \tag{10}$$

Based on weight-redistributed loss, our model can automatically reduce the punishment of easy-classified points and give higher weights on those hard points, thereby enhancing the detection capacity.

4. Experiments

4.1. Implementation details

We use [11] as our baseline to construct a bottom-up multi-person pose estimation system. Our method is trained and evaluated on MS-COCO 2017 dataset [8], which consists of train set (includes 57K images), val set (includes 5K images) and test-dev set (includes 20K images).

Evaluation metric. The average precision and average recall based on Object Keypoint Similarity (OKS) are employed as evaluation metrics. $OKS=\frac{\sum_i\exp(-d_i^2/2s^2k_i^2)\delta(v_i>0)}{\sum_i\delta(v_i>0)}$, where d_i stands for the Euclidean distance between each detected keypoint and its relevant ground truth, v_i shows the visibility flag of ground truth, s represents the object scale, and k_i is a constant which controls falloff. We report the average precision (AP) and recall, including AP (the mean of AP scores at $OKS=0.50,0.55,...,0.90,0.95$), AP^{50} (AP at $OKS=0.50$), AP^{75} (AP at $OKS=0.75$), AP^M for medium scale person and AP^L for large scale person.

Training. We conduct the training process on the COCO train2017 dataset for 140 epochs. The data augmentation technique follows [11] and includes random rotation (randomly from [−30, 30] degrees), random scale (0.75 to 1.25), and random flip horizontally with the probability of 0.5. We crop and resize the training images to 512×512 pixels for our model. We adopt Adam optimizer [28] with the basic learning rate of $1e-3$, and drop it to $1e-4$ and $1e-5$ at the 90th and 120th epochs respectively. We use two GPUs (2080Ti) to train our model and the mini-batch size is 12.

In this work, we set the basic standard deviation σ of Gaussian kernel as 2 and θ_{JNT} as 0.01 according to extensive experiments. The relative scale proportions of all types of keypoints is collected as [0.026, 0.025, 0.025, 0.035, 0.035, 0.079, 0.079, 0.072, 0.072, 0.062, 0.062, 0.107, 0.107, 0.087, 0.087, 0.089, 0.089]. In addition, we set $\alpha_1=0.7$, $\beta_1=-2$, $\alpha_2=-0.5$, $\beta_2=1.75$ and conduct middle supervision between stage3 and stage4 in our model to supervise the training process.

Testing. We resize the short side of input images to 512 pixels and maintain the aspect ratio between width and height. We adopt three scales (0.5, 1, 2) in multi-scale testing and compute the offset maps and heatmaps by resizing all the estimated maps to the same size with average operation.

Table 1
Comparisons with bottom-up methods on the COCO test-dev2017 dataset. * means using refinement.

Method	Input size	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR	AR^{50}	AR^{75}	AR^{M}	AR^{L}
single-scale testing											
Openpose* [18]	368 × 368	61.8	84.9	67.5	57.1	68.2	66.5	87.2	71.8	60.6	74.6
CenterNet-DLA [26]	512 × 512	57.9	84.7	63.1	52.5	67.4					
CenterNet-HG [26]	512 × 512	63.0	86.8	69.6	58.9	70.4					
HrHRNet-W32 [5]	512 × 512	66.4	87.5	72.8	61.2	74.2					
PersonLab [27]	1401 × 1401	66.5	88.0	72.6	62.4	72.3	71.0			66.1	77.7
Pifpaf [22]	641 × 641	66.7			62.4	72.9					
Our Method-W32	512 × 512	66.6	88.0	72.9	61.3	74.3	72.1	91.5	77.5	65.3	81.3
multi-scale testing											
PersonLab [27]	1401 × 1401	68.7	89.0	75.4	64.1	75.5	75.4			69.7	83.0
Our Method-W32	512 × 512	**69.4**	**88.9**	**76.3**	**65.2**	**75.2**	**74.7**	**92.5**	**80.7**	**69.2**	**82.2**

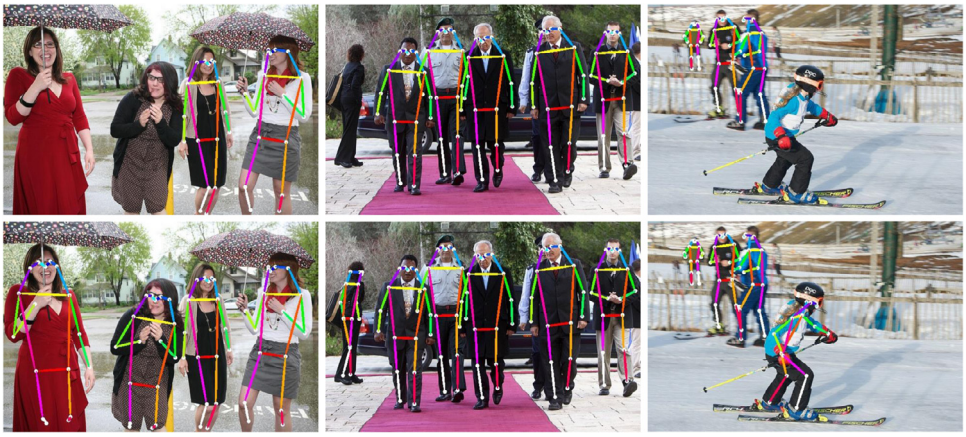

Fig. 6. Visualization of some qualitative results on COCO test2017 dataset. The top row shows the results obtained by the baseline, the bottom row shows the results obtained by our method.

4.2. Results

Table 1 reports the comparisons between our approach with those bottom-up methods on COCO test-dev2017 dataset. We employ HRNet-W32 as the backbone and use three scales for multi-scale testing. To ensure a fair comparison, all the results are reported referring to their original papers.

The results show that our method has achieved 66.6% AP with single-scale tests, which is comparable to those state-of-the-art bottom-up methods (Pifpaf, PersonLab and HrHRNet-W32). If we further adopt a multi-scale test, our method can achieve 69.4% AP, which is highly competitive with other bottom-up methods and narrows the performance gap between top-down and bottom-up methods.

Fig. 6 shows the visualization of some qualitative results, the top row in the figure shows the results obtained by the baseline and the bottom row shows the results obtained by our method. As shown in the figure, there are still some missing cases in the detection results of the baseline, while our model can estimate the pose of each person very well, which reveals our improvement in estimation quality.

4.3. Ablation study

We perform a number of ablation experiments to illustrate the effectiveness of each component in our method. We validate our method on COCO val2017 dataset without multi-scale testing and all the results shown in Table 2 are trained with the same configuration.

Scale-aware heatmap generator. We can see from the Table 2 that the proposed SAHG achieves an improvement of 0.9% AP compared with the baseline, demonstrating the effectiveness of it. We notice that the main gain comes from improving the detection accuracy of large-scale humans, which means the scale variation of keypoints within the human body has a greater impact on those larger-scale people. Besides, since bodies of different scales in the image are based on the same basic variance during the construction of heatmaps, this exacerbates the mutual interference between different keypoints within small-scale bodies and weakens the effect of SAHG on them.

Table 2
Ablation study on the COCO validation dataset. SAHG means scale-aware heatmap generator and WR loss means weight-redistributed loss function.

SAHG	WR loss	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
		65.2	85.8	71.3	60.0	73.6
✓		66.1	85.7	72.2	59.9	75.5
	✓	67.1	86.7	72.9	61.5	75.8
✓	✓	**67.5**	**86.8**	**73.4**	**62.1**	**75.8**

Weight-redistributed loss. Compared with the baseline, our weight-redistributed loss can achieve an improvement of 1.9% AP alone and 2.3% AP together with SAHG, which greatly improves the accuracy of estimation. As we can see in Table 2, model trained with weight-redistributed loss can benefit from both large-scale and medium-scale people, indicating that the imbalance between easy-classified and hard-classified samples widely exists in different scales of people.

Just noticeable threshold (JNT). We perform some experiments to analyse the influence of JNT. The JNT is designed to distinguish foreground pixels from background pixels in the heatmap, so setting its value too high or too low will reduce its effectiveness. In some heatmap based methods[18], pixels with values less than 0.01 in the heatmap are usually ignored in the implementation,

Table 3
Comparison with different JNT on the COCO validation dataset.

θ_{JNT}	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR	AR^{M}	AR^{L}
	65.2	85.8	71.3	60.0	73.6	70.8	64.2	80.3
0.005	66.6	86.6	72.5	61.0	75.3	72.0	65.1	81.8
0.010	**67.5**	**86.8**	**73.4**	**62.1**	**75.8**	**72.8**	**66.2**	**82.1**
0.015	66.9	86.2	73.0	61.2	75.5	72.2	65.2	82.1

Table 4
Comparison with different scale threshold on the COCO validation dataset.

s_{thr}	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR	AR^{M}	AR^{L}
	65.2	85.8	71.3	60.0	73.6	70.8	64.2	80.3
0.035	61.5	83.2	66.4	53.6	73.8	67.5	57.7	81.1
0.062	**66.1**	**85.7**	**72.2**	**59.9**	**75.5**	**71.2**	**63.7**	**81.7**
0.072	65.6	86.0	71.8	59.7	75.0	71.1	63.8	81.5

which somehow fits our definition of JNT, so we explore 0.01 as the initial value of JNT. As we can see in Table 3, the proposed JNT based method achieves better performance compared to the baseline, which shows the effectiveness of our method. When we set θ_{JNT} to 0.01, the trained model performs better than both when it is set to 0.005 and 0.015, indicating that a proper threshold can better improve the performance.

Scale threshold. We further explore the influence of different scale thresholds on the results. The scale threshold s_{thr} is set to limit the minimum scale during the calculation of adaptive variance. As shown in Table 4, there is a large drop in performance when the s_{thr} is set to 0.032, indicating that heatmap with too small response area will affect the network to capture the keypoint-related information. When we set s_{thr} to 0.062 and 0.072, both of them improve the performance over the baseline and s_{thr} with 0.062 performs better than that of 0.072. It reveals that the scale variation of keypoints within human bodies does affect the performance and the proposed SAHG with a suitable threshold could solve this problem effectively.

5. Conclusions

In this paper, we have proposed a SAHG to tackle the scale variation of keypoints in the construction of heatmaps. The generator can produce more precise heatmaps for each type of keypoints based on their relative scales, and it can be readily embedded in data preprocessing for human pose estimation. We also proposed a weight-redistributed loss function to optimize the contribution between different keypoints, in order to further increase the detection accuracy of hard keypoints. Extensive experiments have demonstrated the performance improvement achieved by the proposed method.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 61,871,437 and in part by the Natural Science Foundation of Hubei Province of China under Grant 2019CFA022. The authors would also like to thank the editors and anonymous reviewers for their insightful comments, which greatly improved the quality of this paper.

References

[1] M.A. Fischler, R.A. Elschlager, The representation and matching of pictorial structures, IEEE Trans. Comput. 100 (1) (1973) 67–92.
[2] P.F. Felzenszwalb, D.P. Huttenlocher, Efficient matching of pictorial structures, in: CVPR, 2000, pp. 66–73.
[3] A. Toshev, C. Szegedy, Deeppose: Human pose estimation via deep neural networks, in: CVPR, 2014, pp. 1653–1660.
[4] J.J. Tompson, A. Jain, Y. LeCun, C. Bregler, Joint training of a convolutional network and a graphical model for human pose estimation, in: NIPS, 2014, pp. 1799–1807.
[5] B. Cheng, B. Xiao, J. Wang, H. Shi, T.S. Huang, L. Zhang, Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation, in: CVPR, 2020, pp. 5386–5395.
[6] Y. Cai, Z. Wang, Z. Luo, B. Yin, A. Du, H. Wang, X. Zhou, E. Zhou, X. Zhang, J. Sun, Learning delicate local representations for multi-person pose estimation, arXiv preprint arXiv:2003.04030 (2020).
[7] F. Zhang, X. Zhu, H. Dai, M. Ye, C. Zhu, Distribution-aware coordinate representation for human pose estimation, in: CVPR, 2020, pp. 7093–7102.
[8] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, C.L. Zitnick, Microsoft coco: Common objects in context, in: ECCV, 2014, pp. 740–755.
[9] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollar, Focal loss for dense object detection, in: ICCV, 2017, pp. 2980–2988.
[10] K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, in: CVPR, 2019, pp. 5693–5703.
[11] K. Sun, Z. Geng, D. Meng, B. Xiao, D. Liu, Z. Zhang, J. Wang, Bottom-up human pose estimation by ranking heatmap-guided adaptive keypoint estimates, arXiv preprint arXiv:2006.15480 (2020).
[12] M. Andriluka, L. Pishchulin, P. Gehler, B. Schiele, 2d human pose estimation: New benchmark and state of the art analysis, in: CVPR, 2014, pp. 3686–3693.
[13] L. Ke, M.-C. Chang, H. Qi, S. Lyu, Multi-scale structure-aware network for human pose estimation, in: ECCV, 2018, pp. 713–728.
[14] Y. Bin, X. Cao, X. Chen, Y. Ge, Y. Tai, C. Wang, J. Li, F. Huang, C. Gao, N. Sang, Adversarial semantic data augmentation for human pose estimation, in: ECCV, 2020, pp. 606–622.
[15] W. Tang, P. Yu, Y. Wu, Deeply learned compositional models for human pose estimation, in: ECCV, 2018, pp. 190–206.
[16] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: CVPR, 2017, pp. 2117–2125.
[17] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, J. Sun, Cascaded pyramid network for multi-person pose estimation, in: CVPR, 2018, pp. 7103–7112.
[18] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, in: CVPR, 2017, pp. 7291–7299.
[19] A. Newell, K. Yang, J. Deng, Stacked hourglass networks for human pose estimation, in: ECCV, 2016, pp. 483–499.
[20] W. Li, Z. Wang, B. Yin, Q. Peng, Y. Du, T. Xiao, G. Yu, H. Lu, Y. Wei, J. Sun, Rethinking on multi-stage networks for human pose estimation, arXiv preprint arXiv:1901.00148 (2019).
[21] B. Xiao, H. Wu, Y. Wei, Simple baselines for human pose estimation and tracking, in: ECCV, 2018, pp. 466–481.
[22] S. Kreiss, L. Bertoni, A. Alahi, Pifpaf: Composite fields for human pose estimation, in: CVPR, 2019, pp. 11977–11986.
[23] J. Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, C. Lu, Crowdpose: Efficient crowded scenes pose estimation and a new benchmark, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10863–10872.
[24] J. Li, W. Su, Z. Wang, Simple pose: Rethinking and improving a bottom-up approach for multi-person pose estimation, in: AAAI, 2020, pp. 11354–11361.
[25] Z. Luo, Z. Wang, Y. Huang, L. Wang, T. Tan, E. Zhou, Rethinking the heatmap regression for bottom-up human pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13264–13273.
[26] X. Zhou, D. Wang, P. Krähenbühl, Objects as points, arXiv preprint arXiv:1904.07850 (2019).
[27] G. Papandreou, T. Zhu, L.-C. Chen, S. Gidaris, J. Tompson, K. Murphy, Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model, in: ECCV, 2018, pp. 269–286.
[28] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).