Received: 28 May 2021 | Revised: 3 November 2021 | Accepted: 3 November 2021

IET Image Processing

DOI: 10.1049/ipr2.12404

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH PAPER

A scale-sensitive heatmap representation for multi-person pose estimation

Congju Du | **Han Yu** | **Li Yu**

the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, China

Correspondence
Li Yu, the School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan, 430074, China.
Email: hustlyu@hust.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Number: 61871437; Natural Science Foundation of Hubei Province, Grant/Award Number: 2019CFA022

Abstract

Multi-person pose estimation is a challenging vision task that can be seriously affected by keypoint scale variation. Existing heatmap-based approaches are devoted to reducing the effect by optimizing backbone architecture or loss functions, but the problem of an inaccurate heatmap representation with different keypoint scales still exists. A scale-sensitive heatmap algorithm is presented to generate reasonable spatial and contextual features for the network to predict more precise coordinates, by systematically considering the standard deviation, truncated radius, and shape of Gaussian kernels. Specifically, the scale-sensitive heatmap algorithm contains three parts: inter-person heatmap, limited-area heatmap, and shape-aware heatmap. The inter-person heatmap allocates different standard deviations for each human instance proportionally calculated by the keypoint-based method, the limited-area heatmap defines the truncated radius to limit the influence area of Gaussian kernels, and the shape-aware heatmap modifies the Gaussian kernels generated by some ellipse-shaped joints. Our scale-sensitive heatmap algorithm outperforms the baseline by a considerable margin on the COCO and CrowdPose benchmark datasets. The code and pretrained models are available at https://github.com/ducongju/Scale-sensitive-Heatmap.

1 | INTRODUCTION

Multi-person pose estimation aims to predict the precise locations of all human anatomical keypoints from input images, which is an essential computer vision task. It can be applied in various applications, such as pose tracking [1–3], action recognition [4–6], person re-identification [7, 8] etc.

Early works for pose estimation [9–11] are primarily based on pictorial structures model [12], which represents the human body as a set of joints and the corresponding limb orientations following the human body structure. The major drawback of such methods is the frangibility in complicated poses and the need to design the structure by hand. In recent years, deep convolutional neural networks have made substantial progress in image processing, providing a new branch for pose estimation. DeepPose [13] is the first attempt to apply convolutional networks to handle pose estimation, which trained an AlexNet-like model to regress all the keypoint coordinates directly.

However, this straightforward and intuitive method lacks the surrounding spatial and contextual information, making the network overfitting in training due to the intrinsic visual ambiguity for keypoint localization. The heatmap can be defined as a confidence map generated from the Gaussian function centered at each annotated keypoint, and the heat value indicates the possibility of whether there exists the target keypoint. Correspondingly, we can obtain the joint position prediction by identifying the local maximums in the heatmap [14]. Converting the joint numerical position to the heatmap allows the networks to learn the spatial information around the keypoints, boosting the generalization performance. Since the heatmap representation is more robust than the coordinate representation, most of the mainstream methods leverage the heatmap as the default configuration to encode the location of keypoints.

With the heatmap representation, existing methods for multi-person pose estimation can be classified into top-down methods [15, 16] and bottom-up methods [17, 18]. Top-down

methods first employ a person detector to obtain bounding boxes of people from the input images, following single-person pose estimation. While bottom-up methods directly predict all the body joints and then group them to the corresponding human instance. Scale variation is one of the most challenging problems in multi-person pose estimation, as depicted in Figure 1. The top-down methods can partially solve this problem because the scales of each human joint in the bounding box are almost identical. However, the estimated poses of top-down methods heavily depend on the precision of the person detector, and the computing time is directly affected by the number of persons. On the contrary, the bottom-up methods are efficient (do not need to estimate the pose separately) and concise (end-to-end framework). Hence, the bottom-up methods may have more potential superiority in the wild and crowded scenarios.

On the other hand, it is difficult for the bottom-up methods to simultaneously locate all the keypoints for multiple individuals with various scales. Some bottom-up methods tend to address this problem by optimizing backbone architecture [19] or loss functions [20]. However, the heatmap representation capability is still affected by the scale variation during the process of generating heatmaps. Our goal is to directly integrate the information of keypoint scale variation into heatmaps, hoping that the generated heatmaps can fully represent the spatial and contextual information of the keypoints.

The factors affecting the heatmap representation capability generally include *mean*, *standarddeviation*, *truncatedradius*, and *shape* of Gaussian kernels. The mean of Gaussian kernels is related to the ground truth position, and the other three factors represent the uncertainty of this position. Zhang *et al.* [21] generated the unbiased heatmaps allowing Gaussian kernels to be centered at sub-pixel locations by modifying the *mean* of Gaussian kernels. Their proposed distribution-aware coordinate representation of keypoint method decreases the quantization errors introduced during the resolution reduction. Our works concentrate on reducing the background error caused by the keypoints scale variation and ensuring that heatmaps have adequate spatial and contextual information. Therefore, our scale-sensitive heatmap algorithm focuses on other three factors: *standarddeviation*, *truncatedradius*, and *shape* of Gaussian kernels.

Figure 2 shows that the standard deviation of Gaussian kernels will affect the precision of keypoint localization. As depicted in Figure 2a, we find that the keypoint coordinates in the right person cannot be obtained after pose non-maximum suppression (NMS) operation. Hence, if we set the standard deviation a small value, it makes the network optimization hard and gets unexpected results. When the standard deviation is set to an enormous value, as shown in Figure 2c, keypoints with too much background information will make the network unable to focus on the peak position of the heatmap. In that case, the predicted heatmap may be ambiguous, thus obtaining the inaccurate coordinate by the maximum operation. We can obtain better localization results by setting the proper standard deviation shown in Figure 2b. To address the scale variation problem between human instances, we form a base standard deviation and compute the suitable standard deviation weight for each human instance proportionally, thus incorporating the scale information into heatmaps.

FIGURE 1 Illustration of scale variation situations in multi-person pose estimation. (a) Scale variation caused by occlusion or clothing. (b) Scale variation between keypoints caused by various poses. (c) Scale variation between human instances caused by perspective

Intuitively, if the influence area of Gaussian kernels is not limited, the generated heatmap does not meet the requirements of the human pose estimation task, especially multi-person pose estimation. As shown in Figure 3a, if there is no limit to the area of Gaussian distribution, a pixel far away from the ground

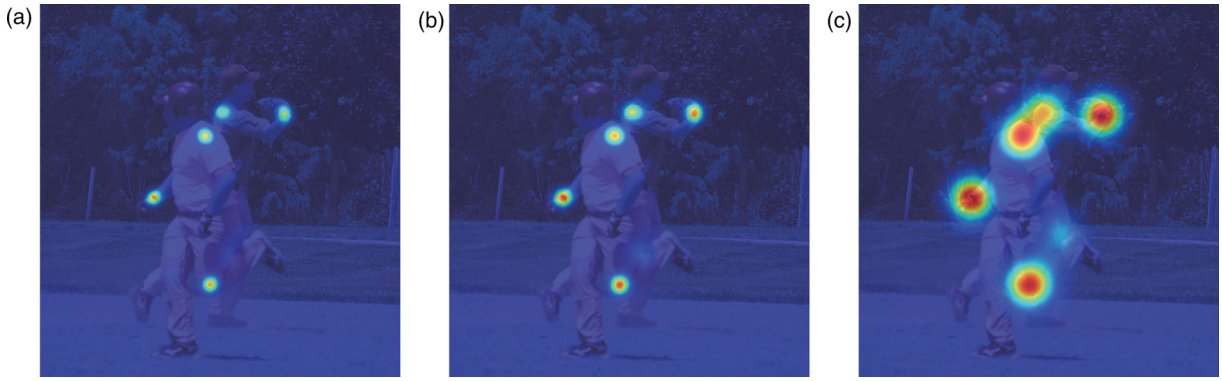

FIGURE 2 Example heatmap outputs predicted by the same backbone network HRNet. There are two people in the image, and we concatenate left shoulder, left knee, and right wrist heatmaps on the input image for illustrating convenience. (a) The Gaussian standard deviation is set to 1. (b) The Gaussian standard deviation is set to 2. (c) The Gaussian standard deviation is set to 5

truth coordinate will still receive a non-zero heat value. This heat value will puzzle the network to learn whether the joint exists in this pixel. Furthermore, the heatmap is computed by superposing each type of keypoints in all human instances, thus making a new peak between the two neighboring Gaussian kernels and reducing the pose estimation performance. To overcome the above problem, we take the truncated radius into consideration to limit the influence area of Gaussian kernels shown in Figure 3b. In the related human keypoints segmentation task, the joints always occupy a particular area and should be annotated as "True" in these pixels. In the same manner, the heat value in the heatmap is annotated as the possibility of whether there exists a joint, and we empirically analyze that the truncated radius of Gaussian kernels should be consistent with the scale of the human joints.

Besides the standard deviation and truncated radius, the shape of Gaussian kernels is the factor affecting the heatmap representation. Some research on anthropometry and human body modeling shows that some joints (such as eyes, ears, and nose) are more ellipse-shaped than round-shaped [12, 22]. Therefore, we want the long axis of these ellipse-shaped joints to represent more spatial information.

Based on the above consideration, we propose a scale-sensitive heatmap algorithm to tackle the scale variation problem caused by perspective. In the method section, we first present the keypoint-based method to compute the relative scale proportions for different human instances and a scale threshold of the person instances to set the lower limit of the standard deviation. The standard deviation will vary with the scale of the person instances. Then, we define the truncated radius in the Gaussian heatmap and calculate it according to the PauTa criterion. When the confidence probability is fixed, the truncated radius is only related to the standard deviation of Gaussian kernels. In that case, we can set the truncated radius adaptively by adjusting the standard deviation. Finally, we generate a shape-aware heatmap for some ellipse-shaped joints by modifying the covariance matrix.

The major contributions of this work can be summarized as follows:

- We propose a scale-sensitive heatmap algorithm to handle the scale variation problem between different human instances, which generates heatmaps for each person according to their relative scale of the person instances;
- We take the *mean*, *standarddeviation*, *truncatedradius*, and *shape* of Gaussian kernels as the essential factors affecting the heatmap representation and point out its impact on keypoint localization. To the best of our knowledge, this is the first attempt to formulate the heatmap representation in heatmap-based pose estimation systematically;
- We propose the keypoint-based method to determine the scale of the person instances and correlate the truncation radius with the standard deviation according to the PauTa criterion.

2 | RELATED WORK

2.1 | Heatmap representation

Compared with regress-based pose estimation methods [13, 23, 24], heatmap-based methods map the joint coordinates to heatmaps, then pose estimation networks are trained by minimizing the discrepancy between the ground truth heatmaps and predicted heatmaps. Tompson et al. [14] firstly proposed to predict heatmaps instead of joints coordinates with the higher-level spatial-model. Inspired by the heatmap-based work, they [25] further extended a coarse-to-fine model. They cropped the convolution features at the localization result of the coarse heatmap and used additional convolutional layers for fine-tuning.

Although heatmaps provide richer contextual and spatial information than joint coordinates, coordinates can be added to network training as auxiliary information. Fan et al. [26] proposed a dual-source convolutional neural network (DS-CNN) which takes image patches and the full-body as inputs to combine the local part appearance and the holistic view of each part. DS-CNN predicted heatmaps together with joint coordinates, and final results are obtained from both. Bulat and Tzimiropoulos [27] designed a cascaded convolutional neural

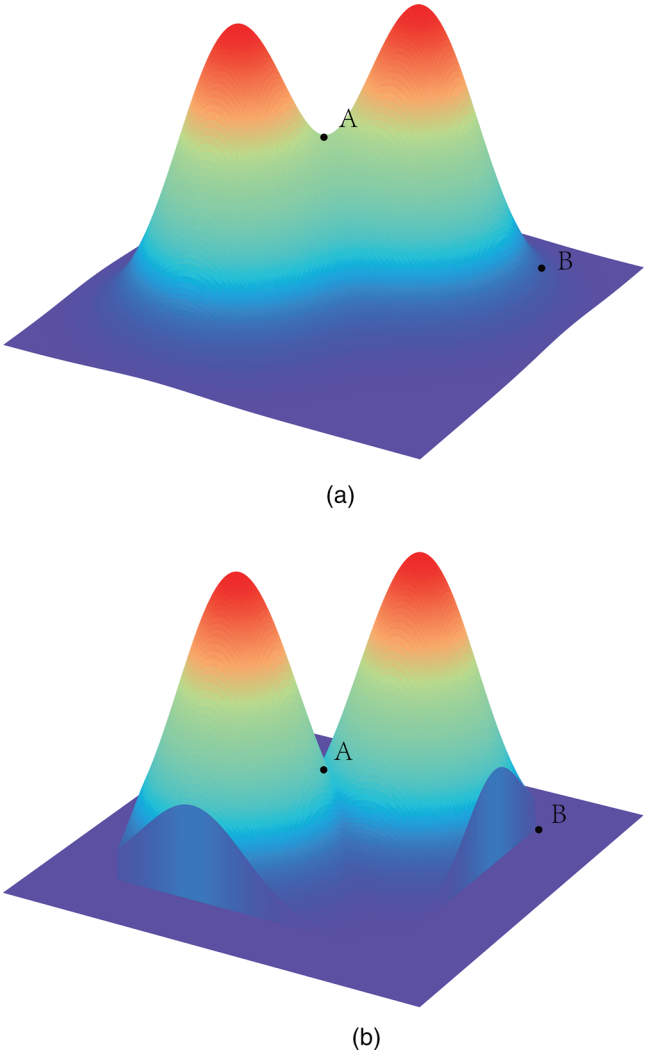

FIGURE 3 Illustration of heatmap aggregation in multi-person pose estimation. The figure contains the same type of keypoints of two people. The x-axis and y-axis represent the pixel position in the heatmap. The z-axis denotes the corresponding heat value. (a) The standard Gaussian heatmap without truncated radius. Two neighboring Gaussian kernels will interfere with each other, making a new peak at point A. Point B shows that pixels far away from all peaks receive an unignorable non-zero heat value. (b) The standard Gaussian heatmap with truncated radius

network with two connected subnetworks. The first branch is a heatmap detection network trained to detect the individual body parts using a per-pixel softmax loss, and the second is performs regression on these heatmaps.

Furthermore, a heatmap representation can be learned indirectly by the regression approach. Luvizon et al. [28] proposed an end-to-end method without heatmap generation, adopting soft-argmax function to convert feature maps into joint coordinates in a fully differentiable framework. At the final stage of the network, the produced feature maps represented joint probabilities at each pixel. Besides the single location pattern information, Tang et al. [29] designed heatmaps with hierarchies of meaningful parts and subparts, called deeply learned compositional model (DLCM). Bone-based heatmap represented the complex and realistic relationships among body parts.

2.2 | Scale variation

Feature pyramid architecture is widely used for settling the scale variation. Lin et al. [30] exploited feature pyramids to extract multi-scale features, designing a pyramidal hierarchy network followed by a number of works [15, 17]. Newell et al. [31] employed a symmetrical network structure to combine different scale representations, which is regarded as a landmark in pose estimation. Xiao et al. [1] utilized a few deconvolutional modules over the last convolution stage in the ResNet, which generated a simple yet strong baseline for high-resolution representations. Sun et al. [16] proposed a novel structure to maintain high-resolution representations over the network, resulting in high performance with the fusion of various scale features. Cheng et al. [19] further leveraged deconvolution module to generate higher resolution representations combining with [1] and [16], achieving state-of-the-art in bottom-up pose estimation.

Additional to the feature pyramid, the loss function is also intended for dealing with the scale variation. Kreiss et al. [32] indicated that localization error of keypoints might be affected by different scales of people in various ways, so it constructed a novel loss function and injected a scale dependence into it. Li et al. [20] observed that plenty of easy-detected samples might attract most of the attention in training, thus impeding the network from learning hard samples. To enhance the discriminating ability of the network, it followed [33] to build a focal L2 loss function to tackle the imbalance problem.

Unlike the work mentioned above, this paper concentrates on solving the challenge of an inaccurate heatmap representation caused by scale variation. We find that it is unreasonable to adopt the same and unconstrained Gaussian kernels to generate heatmaps as different keypoint scales may contain conflicting spatial and contextual information. To this end, we propose a scale-sensitive heatmap algorithm to enhance the network's understanding of the annotated position.

3 | PROPOSED SCALE-SENSITIVE HEATMAP METHOD

The mainstream human pose estimation method is to estimate keypoint heatmaps, followed by choosing the position with the peak values as the keypoints. Compared with them, we formulate reasonable heatmaps to adapt to the scale variation of keypoints. The overall framework of the proposed method is shown in Figure 4. We sort the keypoints from two aspects: inter-person keypoints and ellipse-shaped keypoints. The former means the same types of keypoints within the different human instances, and the latter includes eyes, ears, and nose. The heatmaps of inter-person keypoints are generated by different Gaussian kernels according to their human instance scales. On the basis of inter-person heatmaps, shape-aware heatmaps

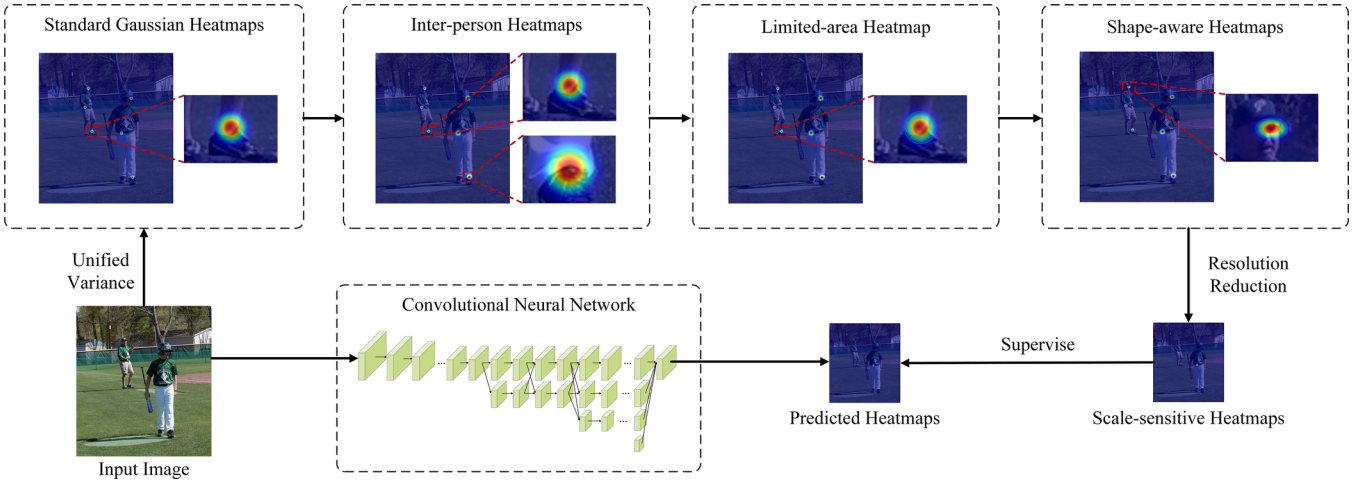

FIGURE 4 A multi-person pose estimation pipeline of the proposed method in the training process. The top row shows the generation of scale-sensitive heatmaps, and we concatenate different types of keypoints' heatmaps on the input image for illustrating convenience. Given an input image, we first obtain adaptive standard deviations for each human instance based on their relative instance scales (called inter-person heatmaps). Secondly, we compute the truncated radius based on the PauTa criterion (called limited-area heatmaps). Finally, we modify the shape of some Gaussian kernels (called shape-aware heatmaps). The bottom row shows the training process supervised by scale-sensitive heatmaps. The final predicted positions can be obtained by the trained CNN and pose NMS, and then human pose can be estimated by pose parsing

modify the Gaussian kernel of some ellipse-shaped keypoints. Therefore, our final scale-sensitive heatmaps can adaptively perceive the scale variation of keypoints in images. In this section, we will describe our scale-sensitive heatmap algorithm in detail.

3.1 | Standard Gaussian heatmaps

To convert keypoint coordinates into heatmaps, most of the existing methods consider each keypoint coordinate as the center of the Gaussian kernel, then superpose all Gaussian kernels generated by the same types of keypoint coordinates. Let $\mathcal{H}^k$ be the ground truth heatmaps, where $k = 1, \ldots, K$ is indexing the keypoint type and K is the number of keypoints. Standard Gaussian heatmaps are often represented as default heatmaps for pose estimation, it adopts unified Gaussian kernels, which can be computed by

$$\mathcal{H}^k_{sg}(x) = \sum_{n=1}^{N} \lambda \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right), \quad (1)$$

where $\lambda = 1/(2\pi|\Sigma|^{\frac{1}{2}})$, x denotes the two-dimensional pixel coordinate, μ represents the Gaussian mean vector (ground truth keypoint coordinate), and N represents the number of human instances. The covariance matrix Σ is set as a diagonal matrix

$$\Sigma = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}, \quad (2)$$

where σ represents the standard deviation of the Gaussian function and the elements on the main diagonal of Σ are equal.

3.2 | Inter-person heatmaps

According to the principles of perspective theory [34], the rays of light from a person that further away enters our eyes at a sharper angle than a person directly in front of us. For the dataset collected by the monocular camera, different human instances may have distinct scales caused by perspective effect [35]. Specifically, the closer the person to the camera, the larger the scale of the human joints, and there is a linear relationship between the scale of the keypoints and person instances. To this end, we first calculate the scale of each person instance to give a reasonable estimate of the geometric distortion, and generate different Gaussian heatmaps for each person accordingly.

In the human pose estimation task, there are two usual references to the scale of the person instances in the ground truth annotation. One is the bounding boxes for human detection, and the other is the masks for human instance segmentation. However, some challenging situations, such as the inclined and complicated pose, may appear in the image (as shown in Figure 5a,d), which is different from the scale of the person instances we expected to describe. We want this scale to be irrelevant with various poses and only related to the somatotype of the human body [36].

Based on the statistics of pose estimation datasets [37–39], the relative distances of eyes, nose, ears, and hips to the center point (the average of the coordinates of all keypoints) are the most stable, which means that these keypoints are less affected by various poses. In order to better represent the scale of the person instances, we choose eyes and hips as the pose-independent keypoints for our method. Therefore, We propose the keypoint-based method using the eyes and hips (illustrated in Figure 5b,e) as a reference instead of bounding boxes or masks (illustrated in Figure 5c,f).

FIGURE 5 Examples of calculating the scale of the person instances in inter-person heatmaps. The top row and the bottom row show the inclined and complicated poses in single-person and multi-person situations, respectively. (a) and (d) Original images. (b) and (e) The illustration of the proposed keypoint-based method, the scale of the person instances is equal to the length of the red line. (c) and (f) The illustration of the contrasting box-based method, the scale of the person instances can be calculated by the area of the bounding box

Suppose that the coordinates of left and right eyes are (x_{le}, y_{le}) and (x_{re}, y_{re}), the coordinates of left and right hips are (x_{lh}, y_{lh}) and (x_{rh}, y_{rh}). We firstly calculate the midpoint coordinates (x_{me}, y_{me}) between the left and right eyes as

$$x_{me} = \frac{x_{le} + x_{re}}{2}, \; y_{me} = \frac{y_{le} + y_{re}}{2}, \tag{3}$$

and the vertical line through the midpoint is used as the symmetrical axis of the human instance. We can get the direction vector $\boldsymbol{s}$ of the symmetric axis by the vector $\boldsymbol{e}$ in the direction of eyes connection

$$\boldsymbol{e} = (x_{le} - x_{re}, y_{le} - y_{re}), \tag{4}$$

$$\boldsymbol{s} = (y_{le} - y_{re}, x_{re} - x_{le}). \tag{5}$$

The distance between (x_{me}, y_{me}) and (x_{lh}, y_{lh}) projected on $\boldsymbol{s}$ is computed as the scale of the person instances

$$\boldsymbol{lh} = (x_{lh} - x_{me}, y_{lh} - y_{me}), \tag{6}$$

$$\left|\boldsymbol{lh}'\right| = \frac{\boldsymbol{s} \cdot \boldsymbol{lh}}{|\boldsymbol{lh}|}. \tag{7}$$

(7) can be simplified according to the ground truth coordinates

$$\left|\boldsymbol{lh}'\right| = \frac{x_{le}(y_{re} - y_{lh}) + x_{re}(y_{lh} - y_{le}) + x_{lh}(y_{le} - y_{re})}{\sqrt{(x_{lh} - x_{me})^2 + (y_{lh} - y_{me})^2}}. \tag{8}$$

The scale of the right ankle $|\boldsymbol{rh}'|$ can be calculated in the same approach. We select the maximum value within $|\boldsymbol{lh}'|$ and $|\boldsymbol{rh}'|$ as the final scale of the person instances.

If we directly establish a linear function of the scale of the person instances and the standard deviation of Gaussian kernels, some small-scale human instances may lack spatial information. We introduce the scale threshold of the person instances to set the lower limit of the truncated radius, forcing some small joints to generate heatmaps with a certain standard deviation. Our inter-person heatmaps can be formulated as

$$\mathcal{H}_{ip}^{k}(x) = \sum_{n=1}^{N} \lambda_n \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma_n^{-1} (x - \mu)\right), \tag{9}$$

and the covariance matrix is set as

$$\Sigma_n = \begin{bmatrix} \sigma_n^2 & 0 \\ 0 & \sigma_n^2 \end{bmatrix}. \tag{10}$$

The diagonal elements of the covariance matrix Σ_n are still equal here, and the standard deviation σ^2 satisfy

$$\sigma_n^2 = \begin{cases} s_n \sigma_{base}^2 / s_{base} & s_n \geq s_{thr} \\ \sigma_{base}^2, & \text{else} \end{cases}, \tag{11}$$

where s_n is the final scale of each person instance, s_{base} is a base scale of the person instances, s_{thr} is the scale threshold of the person instances, meaning that the scale of the small joints below s_{thr} maintain a base standard deviation σ_{base}^2. If the eyes or ankles are not annotated, we treat the scale of the person instances as the base scale and set $\sigma_n^2 = \sigma_{base}^2$. In practice, we use $s_{base} = s_{thr}$.

3.3 | Limited-area heatmaps

There is a particular discrepancy between the representation of heatmaps and keypoint coordinates, as shown in Figure 3. We propose the limited-area heatmaps to settle the problem of imprecise heat value annotation. The manual annotation result is the center coordinate of the joint, and the reasonable truncated radius should be set as close as possible to the edge of the area occupied by the joint in the image. The first and natural idea is to replace the Gaussian distribution with the truncated Gaussian distribution directly

$$\mathcal{H}_{tg}^k(x) = \sum_{n=1}^{N} \frac{\Sigma' \phi\big(\Sigma'(x-\mu)\big)}{\Phi\big(\Sigma'(b-\mu)\big) - \Phi\big(\Sigma'(a-\mu)\big)}, \tag{12}$$

where $\Sigma' = \Sigma^{-1/2}$, $\phi(\cdot)$ is the standard Gaussian distribution with zero mean and unit standard deviation, and $\Phi(\cdot)$ denotes the cumulative distribution function of the standard Gaussian distribution.

The truncated Gaussian distribution limits the influence area of the keypoints in the heatmap through the parameters a and b. However, a and b have no clear physical meaning, and this heatmap representation is computationally intensive. To achieve the same effect, we only need to determine a truncated radius based on the Gaussian distribution, which can be regarded as a distinction between foreground and background pixels. Specifically, the area within the truncated radius meets the Gaussian distribution, and the heat values outside the truncated radius can be directly set to zero. Compared with the Gaussian distribution and truncated Gaussian distribution, heatmaps with the truncated radius cannot guarantee the normalization of the probability density function ($\int \mathcal{H}_{sg}(x)dx = N$ or $\int \mathcal{H}_{tg}(x)dx = N$) after the heatmap encoding. We find that the pixel's heat values within the truncated radius, indicating the possibility of whether the keypoint exists, are not affected by the outside area. That is why we call them heatmaps (or confidence maps) rather than probability maps.

Since the truncated radius is a hyperparameter that needs to be set manually, we propose an indirect method for setting the truncated radius. PauTa criterion can detect outliers under the confidence probability condition, and it is utilized here to determine whether the pixels belong to a joint. After the confidence probability P_c is given, limited-area heatmaps can be obtained by

$$\mathcal{H}_{la}^k(x) = \begin{cases} \mathcal{H}^k(x), & d(x,\mu) \leq r_{la} \\ 0, & \text{else}, \end{cases} \tag{13}$$

$$\text{s.t. } \mathrm{P}\left(d(x,\mu) \leq r_{la}\right) = \mathrm{P}_c, \tag{14}$$

where $\mathcal{H}^k(x)$ represents the heatmaps without the truncated radius, $d(p,q) = \max(|p_x - q_x|, |p_y - q_y|)$ is the Chebyshev distance, r_{la} is the truncated radius that limits the area of the heatmap. Given the confidence probability $\mathrm{P}_c = \mathrm{erf}(\alpha/\sqrt{2})$, we can get the truncated radius as α times the standard deviation. The truncation radius goes to infinity when $\mathrm{P}_c = 100\%$.

3.4 | Shape-aware heatmaps

According to the facial anatomy [40, 41], the vertical length of the eye is usually shorter than the widest distance between temporal and nasal sides of the eye, causing oblate in shape for most eyes. Similarly, anatomy and anthropometry [42] systematically measured the height and width of different types of keypoints. They revealed that some facial keypoints like eyes, ears, and nose are more ellipse-shaped than round-shaped. As the relative positions between the facial keypoints are more stable, we can generate more accurate heatmaps to fit the influence area of the joints. Therefore, we propose to modify the shape of heatmaps for these elliptical-shaped keypoints. Fortunately, it is effortless to generate shape-aware Gaussian heatmaps by modifying the covariance matrix of the ellipse-shaped keypoints.

For the convenience of analysis, we take eyes and ears as examples. As the human skeleton model is roughly symmetrical, we find that the major axis of an ellipse is parallel to the line connecting the left and right eyes, and the minor axis is perpendicular to the line. Then we can determine the covariance matrix of the Gaussian kernel according to the angle θ between the vector of eyes connection and the horizontal vector. The direction vector of the eye connection is defined from the left to the right eye. In contrast, the major axis of the ears is perpendicular to the line connecting the left and right ears. Finally, our shape-aware heatmaps can be obtained by

$$\mathcal{H}_{sa}^k(x) = \sum_{n=1}^{N} \lambda_{kn} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma_{kn}^{-1}(x-\mu)\right) \mathcal{M}(\theta), \tag{15}$$

where $\lambda_{kn} = 1/(2\pi|\Sigma_{kn}|^{\frac{1}{2}})$, $\mathcal{M}(\theta)$ is a rotation matrix used to perform the rotation operation in the heatmap

$$\mathcal{M}(\theta) = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}, \tag{16}$$

ALGORITHM 1 Scale-sensitive Heatmap Algorithm

Input: An RGB image I and corresponding ground truth keypoint coordinate (x_{nk}, y_{nk})

Output: Scale-sensitive heatmaps $\mathcal{H}_{ss}^{k}$

```
for n = 1 to N do
    Calculate the scale of each person instance s_n in (7);
    Calculate the standard deviation of each person instance σ_n in (11);
    for k = 1 to K do
        if keypoint not in (eyes, ears, nose) then
            Calculate the covariance matrix Σ_k in (10);
        else
            Calculate the rotation matrix M(θ) in (16);
            Calculate the covariance matrix Σ_k in (17);
        end if
    end for
    Calculate the truncated radius r_la in (14);
    Generate scale-sensitive heatmaps H_ss^k in (13) and (15);
end for
```

and the covariance matrix is modified to

$$\Sigma_{kn} = \begin{bmatrix} \sigma_n^2 & 0 \\ 0 & \omega_k \sigma_n^2 \end{bmatrix}, \tag{17}$$

where ω_k denotes the ratio of minor axis to major axis for eyes, and ears are the opposite. If k not belongs to ellipse-shaped joint, $\theta = 0$ in (16) and $\omega_k = 1$ in (17).

By taking *standarddeviation*, *truncatedradius*, and *shape* of Gaussian kernels into consideration, our scale-sensitive heatmaps include both inter-person heatmaps, limited-area heatmaps, and shape-aware heatmaps. For clarity, the main steps of the scale-sensitive heatmap algorithm are shown in Algorithm 1.

4 | EXPERIMENTAL RESULTS

4.1 | Dataset and evaluation

We perform experiments on two challenging datasets: Microsoft COCO [38] and CrowdPose [39] to demonstrate the effectiveness of our scale-sensitive heatmap algorithm.

Microsoft COCO is a large-scale object detection, segmentation, and keypoint detection dataset, including 250k human instance annotations with 17 keypoints. COCO dataset consists of train2017 set (includes 57k images), val2017 set (includes 5K images), and test-dev2017 set (includes 20K images), covering a wide variety of human poses. CrowdPose is a dataset of crowded human poses, which consists of 20k images and 80k human instances. The training, validation, and testing subsets are split in proportion to 5:1:4. It covers crowded scenes and daily life scenes, encouraging networks to adapt to different kinds of cases.

For COCO and CrowdPose datasets, Average Precision (AP) and Average Recall (AR) based on Object Keypoint Similarity (OKS) are used to evaluate the results. $\text{OKS} = \frac{\sum_i \exp(-d_i^2/2s^2k_i^2)\delta(v_i>0)}{\sum_i \delta(v_i>0)}$, where d_i are the Euclidean distance between each detected keypoint and its relevant ground truth, v_i stands for the visibility flag of ground truth, s represents the object scale, and k_i is a constant that controls falloff. The OKS is a similarity measure between ground truth keypoints and predicted keypoints, which plays the same role as the Intersection over Union (IoU) in object detection. We report AP (the means of AP scores at $\text{OKS} = 0.50, 0.55, \ldots, 0.90, 0.95$), AP^{50} (AP at $\text{OKS} = 0.50$), AP^{75} (AP at $\text{OKS} = 0.75$) as standard metrics. Besides, AP^{M} and AP^{L} is for medium and large scale persons in the COCO dataset. The CrowdPose dataset are divided into three crowding levels by crowd index, for details see [39]. AP^{E}, AP^{M}, and AP^{H} are used in the CrowdPose dataset standing for easy, medium and hard instances respectively.

4.2 | Implementation details

We adopt the pose estimation network HRNet-W32 [16] and HrHRNet-W32 [19] as backbones and use [45] to group the predicted keypoints into different people. Our data argumentation includes random rotation ([−30, 30]), random scale ([0.75, 1.5]), and and random translation ([-40, 40]). We crop and resize the training images to 512×512. For COCO/CrowdPose datasets, the models are trained with Adam optimizer [50] for 140/300 epochs with a mini-batch of 64 images. The initial learning rate is set as 1e-3 and dropped to 1e-4 and 1e-5 after 90/200 and 120/260 epochs. We conduct our training experiments on four NVIDIA Tesla V100 GPUs and implement them by using PyTorch [51] deep learning framework.

During testing, we first resize the short side of the images to 512 and maintain the aspect ratio between height and width. We adopt multi-scale testing with three scales (0.5, 1, 2) and compute heatmaps and offset maps by averaging to the same size. Flip testing is used for all the experiments.

4.3 | Comparisons with state-of-the-arts

To demonstrate the effectiveness of our method, we first compare our multi-person pose estimation result with the previous state-of-the-art method on the COCO dataset. All the results of other methods are obtained referring to their original papers for fair comparisons.

Tables 1 and 2 report the input size and quantitative results compared with some state-of-the-art approaches on the COCO test-dev2017 set. Our method achieves 70.0 AP with multi-scale testing and 68.4 AP with single-scale testing, which is competitive to other state-of-the-art bottom-up methods, even better than PersonLab [18] with high-resolution input images (1401×1401). For a fair comparison, we use HRNet-W32 and HrHRNet-W32 as backbone respectively, and compare them with the same configuration of SOTA methods HRNetBU [45] and SWAHR [46]. Our scale-sensitive heatmap algorithm outperforms HRNetBU/SWAHR by +1.8/+0.5 AP in single-

TABLE 1 Comparisons of AP and AR results with single-scale testing on the COCO test-dev2017 dataset

Method	Backbone	Input size	Refine	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR	AR^{50}	AR^{75}	AR^{M}	AR^{L}
Openpose [17]	CPM	368 × 368	N	52.9	-	-	50.9	57.2	57.0	79.2	-	-	-
Openpose [17]	CPM	368 × 368	Y	61.8	84.9	67.5	57.1	68.2	66.5	87.2	71.8	60.6	74.6
AE [43]	Hourglass 4-stack	512 × 512	N	56.6	81.8	61.8	49.8	67.0	-	-	-	-	-
AE [43]	Hourglass 4-stack	512 × 512	Y	62.8	84.6	69.2	57.5	70.6	-	-	-	-	-
CenterNet [44]	DLA-34	512 × 512	N	57.9	84.7	63.1	52.5	67.4	-	-	-	-	-
CenterNet [44]	Hourglass-104	512 × 512	N	63.0	86.8	69.6	58.9	70.4	-	-	-	-	-
PersonLab [18]	ResNet-101	1401 × 1401	N	65.5	87.1	71.4	61.3	71.5	70.1	89.7	75.7	65.0	77.1
PersonLab [18]	ResNet-152	1401 × 1401	N	66.5	88.0	72.6	62.4	72.3	71.0	90.3	76.6	66.1	77.7
HrHRNet [19]	HrHRNet-W32	512 × 512	N	66.4	87.5	72.8	61.2	74.2	-	-	-	-	-
HRNetBU [45]	HRNet-W32	512 × 512	N	66.6	87.8	72.8	61.1	74.5	71.4	90.8	77.0	64.6	80.8
SWAHR [46]	HrHRNet-W32	512 × 512	N	67.9	88.9	74.5	62.4	75.5	-	-	-	-	-
Ours	HRNet-W32	512 × 512	N	67.5	88.2	73.7	62.1	75.2	72.4	91.5	78.0	65.9	81.4
Ours	HrHRNet-W32	512 × 512	N	**68.4**	**88.7**	**74.5**	**62.2**	**75.5**	**73.0**	**92.1**	**79.0**	**66.1**	**82.2**

The bold values indicate the best results of the listed experimental group. "Refine" indicates whether or not the result is additionally refined by a single-person pose model.

TABLE 2 Comparisons of AP and AR results with multi-scale testing on the COCO test-dev2017 dataset

Method	Backbone	Input size	Refine	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR	AR^{50}	AR^{75}	AR^{M}	AR^{L}
AE [43]	Hourglass 4-stack	512 × 512	N	63.0	85.7	68.9	58.0	70.4	-	-	-	-	-
AE [43]	Hourglass 4-stack	512 × 512	Y	65.5	86.8	72.3	60.6	72.6	70.2	89.5	76.0	64.6	78.1
HGG [47]	Hourglass 4-stack	512 × 512	N	67.6	85.1	73.7	62.7	74.6	71.3	-	-	-	-
PersonLab [18]	ResNet-101	1401 × 1401	N	67.8	88.6	74.4	63.0	74.8	74.5	92.2	80.4	68.6	82.5
PersonLab [18]	ResNet-152	1401 × 1401	N	68.7	89.0	75.4	64.1	75.5	**75.4**	92.7	**81.2**	**69.7**	83.0
HRNetBU [45]	HRNet-W32	512 × 512	N	69.4	88.9	76.2	64.9	75.8	74.9	92.8	81.0	69.1	82.9
Ours	HRNet-W32	512 × 512	N	69.6	89.0	76.2	65.1	76.0	75.0	**93.1**	80.9	69.3	82.9
Ours	HrHRNet-W32	512 × 512	N	**70.0**	**89.2**	**77.1**	**65.6**	**76.4**	75.2	93.0	**81.3**	69.4	**83.0**

scale testing. Compared with HRNetBU method, our multi-scale testing result brings 0.2 AP gain with the same HRNet-W32 backbone. We consider that the aggregation of the results after rescaling the image can alleviate the scale variation problem to some extent, so the multi-scale testing provides a slight improvement.

The results of our method and other state-of-the-art approaches on COCO val2017 set are presented in Table 3. We see that the proposed scale-sensitive heatmap algorithm achieves an overall 68.3 AP, which is still higher than the others. If we further use multi-scale testing, our AP score can exceed 70.0 without any refinement. And if with HrHRNet-W32 backbone, it can finally achieve 71.2 AP, leading to 4.1 AP gain over HrHRNet.

To better evaluate the performance in different crowded scenarios, we also train our models on the CrowdPose training and validation datasets and display the results on the test dataset. Quantitative results, including three crowding levels, are summarized in Table 4. As the CrowdPose dataset includes more scale variation situations, our method brings more remarkable performance improvement than the COCO dataset. Without bells and whistles, our method outperforms the top-down method SimpleBaseline [1] by a large margin of 7.2 AP. Top-down approaches usually fail in crowded scenes because the single-person bounding box may contain erroneous keypoints from other human instances. In addition, our performance surpasses the bottom-up HrHRNet-W48 model even with the HRNet-W32 backbone.

4.4 | Ablation experiments

In this section, we further perform the ablation experiments for the three components in our scale-sensitive heatmaps: inter-person heatmaps, limited-area heatmaps, and shape-aware heatmaps. For fair comparisons, we use HRNet-W32 as the backbone network and take the results of the standard Gaussian heatmaps as our baseline.

4.4.1 | Inter-person heatmaps

We perform an ablation study comparing box-based and keypoint-based methods shown in Table 5. We apply the trun-

TABLE 3 Comparisons of AP and AR results with on the COCO val2017 dataset

Method	Backbone	Input size	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR	AR^{50}	AR^{75}	AR^{M}	AR^{L}
CenterNet [44]	DLA-34	512 × 512	58.9	-	-	-	-	-	-	-	-	-
CenterNet [44]	Hourglass-104	512 × 512	64.0	-	-	-	-	-	-	-	-	-
PersonLab [18]	ResNet-152	601 × 601	54.1	76.4	57.7	40.6	73.3	57.7	-	-	43.5	77.4
PersonLab [18]	ResNet-152	1401 × 1401	66.5	86.2	71.9	62.3	73.2	70.7	-	-	65.6	77.9
HrHRNet [19]	HrHRNet-W32	512 × 512	67.1	86.2	73.0	61.5	76.1	-	-	-	-	-
HRNetBU [45]	HRNet-W32	512 × 512	67.8	86.8	74.0	62.0	76.4	72.3	89.8	77.6	65.6	82.0
Ours	HRNet-W32	512 × 512	68.3	87.2	74.4	62.9	76.5	73.1	90.6	78.2	66.7	82.2
Ours	HrHRNet-W32	512 × 512	69.3	87.3	75.5	64.0	77.2	74.3	91.2	79.9	68.0	83.4
Ours+	HRNet-W32	512 × 512	70.6	**88.0**	76.8	66.1	77.6	75.9	92.2	81.4	70.2	84.1
Ours+	HrHRNet-W32	512 × 512	**71.2**	87.7	**77.4**	**66.9**	**77.9**	**76.5**	**92.2**	**81.9**	**71.1**	**84.3**

+means using multi-scale test.

TABLE 4 Comparisons of AP results on CrowdPose test dataset

Method	AP	AP^{50}	AP^{75}	AP^{E}	AP^{M}	AP^{H}
Openpose [17]	-	-	-	62.7	48.7	32.3
Mask-RCNN [48]	57.2	83.5	60.3	69.4	57.9	42.0
SimpleBaseline [1]	60.8	81.4	65.7	71.4	61.2	51.2
RMPE [49]	61.0	81.3	66.0	71.2	61.4	51.1
HRNetBU [45]	64.9	84.5	69.6	72.7	65.5	56.1
HrHRNet-W48 [19]	65.9	86.4	70.6	73.3	66.5	57.9
Our Method-W32	66.2	85.1	71.4	73.5	66.7	58.0
Our Method-W32+	**68.0**	**86.2**	**73.2**	**75.7**	**68.8**	**59.4**

cated radius to Gaussian kernels, and the confidence probability $P_c = 99.7\%$. As can be seen, both methods exceed the baseline, and our keypoint-based method is better than the box-based method. It demonstrates that the keypoint-based method is more reasonable for describing the scale of the person instances. If the scale threshold of the person instances s_{thr} is set to 256 (our input size is 512 × 512), the best result achieves 67.8 AP over the baseline, proving the better representation of the inter-person heatmaps.

4.4.2 | Limited-area heatmaps

To limit the influence area of Gaussian kernels, we consider the truncated radius into the inter-person heatmaps. Here, several different truncated radii are tested for comparison in Table 6. We choose the keypoint-based method and $s_{thr} = 256$ as our configuration of inter-person heatmaps. $P_c = 100\%$ means that we do not apply the truncated radius. The results show that limited-area heatmaps with $P_c = 99.7\%$ (which is equivalent to $r_{la} = 3\sigma$) get the best pose estimation performance. Furthermore, compared with the heatmaps without the truncated radius, limited-area heatmaps speed up the training owing to the less computational cost of the heatmap generation.

TABLE 6 Ablation experiments of limited-area heatmaps on the COCO val2017 dataset

P_c	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR	AR^{50}	AR^{75}	AR^{M}	AR^{L}
100%	67.5	86.8	73.5	61.4	76.7	72.3	90.0	77.5	65.1	82.4
99.7%	**67.8**	**87.0**	**74.2**	**61.8**	**77.1**	**72.7**	**90.3**	**78.1**	**65.7**	**82.8**
95.5%	67.3	86.4	73.0	61.3	76.5	72.1	89.6	77.0	65.1	82.1

TABLE 7 Ablation experiments of shape-aware heatmaps on the COCO val2017 dataset

ω_k	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR	AR^{50}	AR^{75}	AR^{M}	AR^{L}
1:1	66.3	85.8	72.3	59.8	76.4	71.5	89.3	76.4	63.9	82.3
16:9	**67.6**	**86.7**	73.5	**61.3**	**77.0**	**72.3**	89.7	77.4	**65.2**	**82.5**
2:1	67.4	86.6	**73.8**	61.3	76.6	72.2	**89.9**	**77.6**	65.2	82.1

TABLE 5 Ablation experiments of inter-person heatmaps on the COCO val2017 dataset

Method	s_{thr}	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR	AR^{50}	AR^{75}	AR^{M}	AR^{L}
Baseline	-	66.3	85.8	72.3	59.8	76.4	71.5	89.3	76.4	63.9	82.3
Box-based	192	67.2	86.8	73.1	61.1	76.5	72.2	90.2	77.1	65.2	82.2
Box-based	256	67.4	86.4	73.2	61.0	76.9	72.2	89.7	77.2	65.1	82.4
Keypoint-based	192	67.7	87.0	73.3	**61.5**	**77.4**	72.6	90.1	**78.0**	**65.6**	82.7
Keypoint-based	256	**67.8**	**87.1**	**73.9**	**61.5**	**77.4**	**72.7**	**90.2**	77.9	65.5	**83.0**

TABLE 8 Ablation experiments of scale-sensitive heatmaps on the COCO val2017 dataset

	Inter-person	Limited-area	Shape-aware	AP	AP^{50}	AP^{75}	AP^M	AP^L	AR	AR^{50}	AR^{75}	AR^M	AR^L
(a)				66.3	85.8	72.3	59.8	76.4	71.5	89.3	76.4	63.9	82.3
(b)	✓			67.5	86.8	73.5	61.4	76.7	72.2	90.2	77.1	65.2	82.2
(c)	✓	✓		67.8	87.0	74.2	61.8	**77.1**	72.7	90.3	78.1	65.7	**82.8**
(d)		✓	✓	67.6	86.7	73.5	61.3	77.0	72.3	89.7	77.4	65.2	82.5
(e)	✓	✓	✓	**68.3**	**87.2**	**74.4**	**62.9**	76.5	**73.1**	**90.6**	**78.2**	**66.7**	82.2

TABLE 9 Comparison with different standard deviations of on the MPII test dataset

σ	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
1	96.5	95.5	89.6	84.5	88.4	86.0	81.2	89.3
2	**97.3**	**95.9**	**90.4**	**85.9**	**88.7**	**86.5**	**82.6**	**90.1**
5	97.0	95.7	89.6	84.9	89.2	85.0	80.5	89.3

TABLE 10 Comparison with different standard deviations of on the COCO val2017 dataset

σ	AP	AP^{50}	AP^{75}	AP^M	AP^L
1	65.3	85.1	71.0	58.3	76.3
2	**66.3**	**85.8**	**72.3**	**59.8**	**76.4**
5	58.5	80.4	63.2	48.4	73.9

TABLE 11 Comparison with different standard deviations of on CrowdPose test dataset

σ	AP	AP^{50}	AP^{75}	AP^E	AP^M	AP^H
1	64.6	83.7	69.5	72.0	65.3	55.9
2	**66.2**	**85.1**	**71.4**	**73.5**	**66.7**	**58.0**
5	58.6	80.9	62.2	67.0	58.8	50.8

4.4.3 | Shape-aware heatmaps

We perform an ablation study on the effect of the ratio parameter ω_k. As shown in Table 7, $\omega_k = 1:1$ means without shape-aware heatmaps, $\omega_k = 16:9$ achieves the best 67.6 AP. It proves that Gaussian kernels with a shape closer to the ground truth keypoints can encourage the network to predict more accurate locations.

4.4.4 | Scale-sensitive heatmaps

Our complete framework combines inter-person heatmaps, limited-area heatmaps, and shape-aware heatmaps to get the final scale-sensitive heatmaps. Table 8 reports the ablation experiments on the COCO val2017 dataset. Comparing method (a) and (d), shape-aware heatmaps with the truncated radius achieves 67.6 AP, which is 1.3 AP better than the baseline. By comparing method (c) and (e), we can find that the performance on large person will be degraded from 77.1 to 76.5 with the shape-aware modification. We consider that large-scale human joints contain rich feature representations that obscure the shape information. Adding shape-aware modification to the Gaussian kernel tends to break structural consistency and makes the network difficult to train, which may degrade its performance.

We analyze that the inherent ambiguity of Gaussian kernels limits the standard deviation, and Gaussian kernels with large standard deviations are more likely to overlap in crowded scenes. Our method aims to make the influence area of Gaussian kernels closer to the scale of the human joints, then we allocate different standard deviations for each human instance proportionally. Therefore, our improvements mainly come from medium persons with medium standard deviations. Specifically, the final result (method (e)) is further improved to 68.3 AP (+2.0 AP), and we get a large gain (+3.1 AP^M and +2.8AR^M) in medium person poses. There is a slight drop in the recall of large person (-0.1 AR^L), and the precision for large person poses no reduction (+0.1 AP^L). It demonstrates that our scale-sensitive heatmaps boost the ability to perceive small and medium person poses while ensuring the localization accuracy of large-scale persons.

4.5 | Further discussions

4.5.1 | Standard deviation

We consider that the standard deviation of Gaussian kernels will affect the pose estimation performance during heatmap generation. To verify this, we conducted experiments on three datasets: COCO, CrowdPose, and MPII.

MPII Human Pose dataset [37] includes around 25K images containing over 40K annotated poses, where roughly three-quarters of the collected images are used for training, and the rest are used for testing. The images are systematically extracted from YouTube videos covering everyday human activities with full-body human poses. For the MPII dataset, we use the standard Percentage of Correct Keypoints (PCK) metric defined in [52], which measures the percentage of the joint position that falls within a normalized distance of the ground truth. A slight modification of the "PCK" uses the matching threshold as 50% of the head segment length, called "PCKh@0.5".

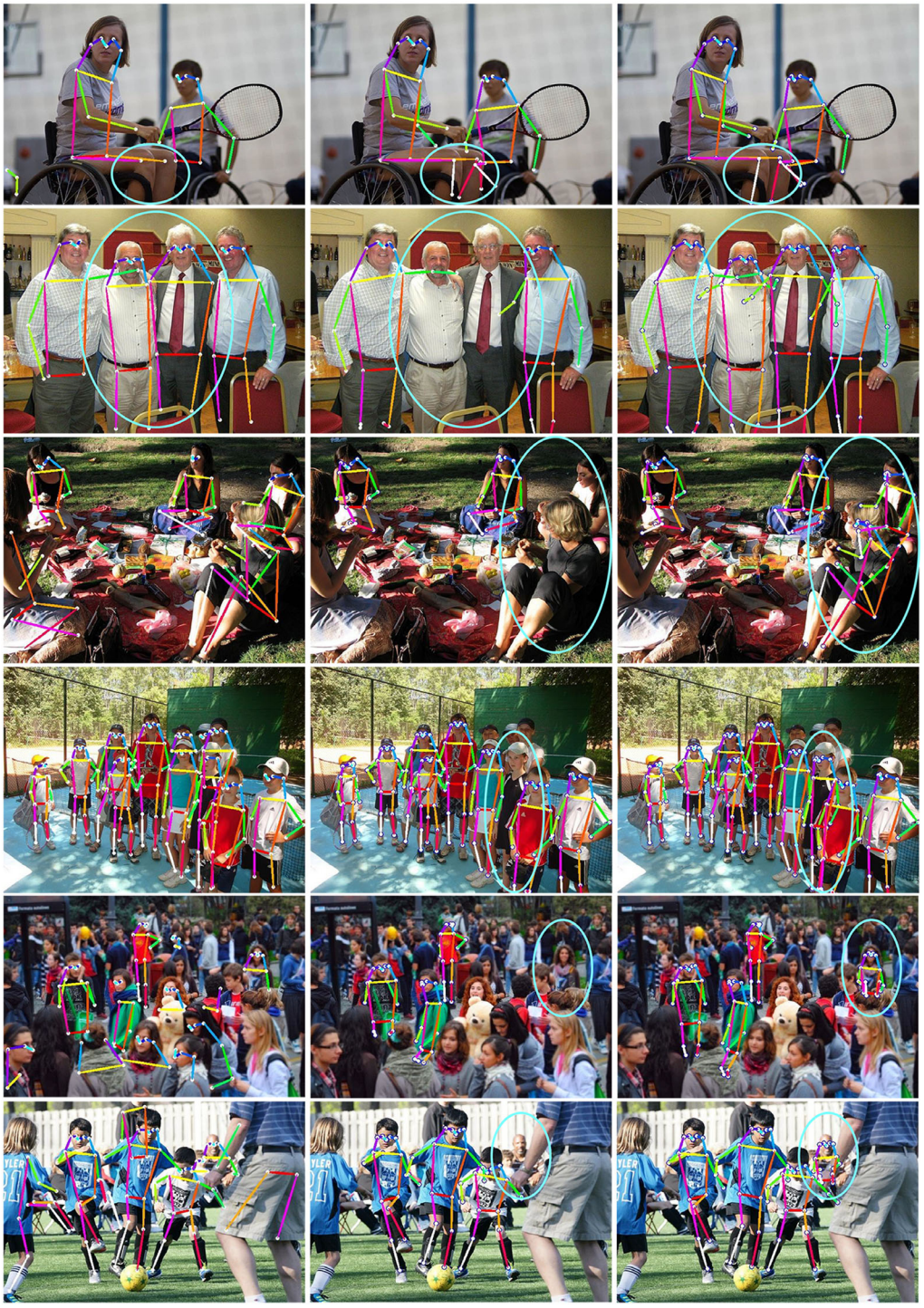

FIGURE 6 Comparison of qualitative results on the COCO dataset with occlusion and scale variation situations. (a) Ground truth. (b) Results by baseline method. Results by our method

Table 9 shows the PCKh results with the standard deviations of 1, 2, and 5 on the MPII dataset. Tables 10 and 11 show the AP/AR results with the same configuration on the COCO and CrowdPose datasets. Results on three datasets prove the point we mentioned in Section 1: excessively small ($\sigma = 1$) or large ($\sigma = 5$) standard deviation reduces the pose estimation performance. The heatmap generated by an appropriate standard deviation allows the network to extract spatial and contextual information near the joint position for better inference and eliminate some useless background pixels simultaneously.

4.5.2 | Qualitative discussion

Figure 6 illustrates some qualitative results for comparison on the COCO val2017 dataset. The first and second rows show the prediction under the occlusion scenarios. Our method predicts more true positives than baseline and even reasonably infers the occluded keypoints that are missing annotated in the ground truth. The third to sixth rows give the qualitative comparison under the scale variation situations. They show better results on large-scale (the third row), medium-scale (the fourth and fifth rows), and small-scale (the sixth row) human instances compared to baseline, demonstrating the effectiveness of our method on scale variation problems. Some keypoints in crowded scenarios are still not predicted, probably because the ground truth is not exhaustively annotated with all the keypoints (see the fifth row).

5 | CONCLUSION

In this paper, we present a scale-sensitive heatmap algorithm for multi-person pose estimation. The main idea is to limit the influence area of Gaussian distribution by generating heatmaps based on the corresponding scale of keypoints. The inter-person heatmap calculates the scale of the person instances by the keypoint-based method, and the limited-area heatmap sets the truncated radius accordingly. The shape-aware heatmap revises the shape of Gaussian kernels to represent more accurate spatial and contextual information. The qualitative and quantitative results show that our method achieves a better performance in multi-person pose estimation, especially for small and medium persons. Our scale-sensitive heatmap method is a heuristic work, not merely for human pose estimation. Future works will focus on other computer vision tasks about heatmaps, such as object tracking or action recognition, which utilize each heat value instead of just the local maximum in the heatmap.

ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant 61871437 and in part by the Natural Science Foundation of Hubei Province of China under Grant 2019CFA022. The computing work in this paper is supported by the public computing service platform provided by the Network and Computing Center of HUST.

The authors would also like to thank the editors and anonymous reviewers for their insightful comments, which greatly improved the quality of this paper.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in "Scale-sensitive-Heatmap" at https://github.com/ducongju/Scale-sensitive-Heatmap.

ORCID

Congju Du https://orcid.org/0000-0002-1274-319X
Han Yu https://orcid.org/0000-0003-1858-1758
Li Yu https://orcid.org/0000-0002-5060-2558

REFERENCES

1. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 466–481. Springer, Berlin (2018)
2. Ruan, W., Liu, W., Bao, Q., Chen, J., Cheng, Y., Mei, T.: Poinet: pose-guided ovonic insight network for multi-person pose tracking. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 284–292. ACM, New York (2019)
3. Raaj, Y., Idrees, H., Hidalgo, G., Sheikh, Y.: Efficient online multi-person 2d pose tracking with recurrent spatio-temporal affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4620–4628. IEEE, Piscataway (2019)
4. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). vol. 32. AAAI Press, Palo Alto (2018)
5. Ghazal, S., Khan, U.S., Saleem, M.M., Rashid, N., Iqbal, J.: Human activity recognition using 2d skeleton data and supervised machine learning. IET Image Proc. 13(13), 2572–2578 (2019)
6. Luvizon, D., Picard, D., Tabia, H.: Multi-task deep learning for real-time 3d human pose estimation and action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020)
7. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3908–3916. IEEE, Piscataway (2015)
8. Zhao, Y.B., Lin, J.W., Xuan, Q., Xi, X.: Hpiln: a feature learning framework for cross-modality person re-identification. IET Image Proc. 13(14), 2897–2904 (2019)
9. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3d human pose annotations. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1365–1372. IEEE, Piscataway (2009)
10. Sun, M., Savarese, S.: Articulated part-based model for joint object detection and pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 723–730. IEEE, Piscataway (2011)
11. Sapp, B., Taskar, B.: Modec: Multimodal decomposable models for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3674–3681. IEEE, Piscataway (2013)
12. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. Int. J. Comput. Vision 61(1), 55–79 (2005)
13. Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1653–1660. IEEE, Piscataway (2014)
14. Tompson, J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. In: Proceedings of the International Conference on Neural Information Processing Systems (NIPS), pp. 1799–1807. MIT Press, Cambridge (2014)
15. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7103–7112. IEEE, Piscataway (2018)
16. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5693–5703. IEEE, Piscataway (2019)
17. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE

Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7291–7299. IEEE, Piscataway (2017)
18. Papandreou, G., Zhu, T., Chen, L.C., Gidaris, S., Tompson, J., Murphy, K.: Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 269–286. IEEE, Piscataway (2018)
19. Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T.S., Zhang, L.: Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5386–5395. IEEE, Piscataway (2020)
20. Li, J., Su, W., Wang, Z.: Simple pose: Rethinking and improving a bottom-up approach for multi-person pose estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), pp. 11354–11361. AAAI Press, Palo Alto (2020)
21. Zhang, F., Zhu, X., Dai, H., Ye, M., Zhu, C.: Distribution-aware coordinate representation for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7093–7102. IEEE, Piscataway (2020)
22. Wang, Y., Mori, G.: Multiple tree models for occlusion and spatial constraints in human pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 710–724. (2008)
23. Belagiannis, V., Rupprecht, C., Carneiro, G., Navab, N.: Robust optimization for deep regression. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2830–2838. IEEE, Piscataway (2015)
24. 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 (CVPR), pp. 4733–4742. IEEE, Piscataway (2016)
25. 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 (CVPR), pp. 648–656. IEEE, Piscataway (2015)
26. Fan, X., Zheng, K., Lin, Y., Wang, S.: Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1347–1355. IEEE, Piscataway (2015)
27. Bulat, A., Tzimiropoulos, G.: Human pose estimation via convolutional part heatmap regression. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 717–732. Springer, Berlin (2016)
28. Luvizon, D.C., Tabia, H., Picard, D.: Human pose regression by combining indirect part detection and contextual information. Computers & Graphics 85, 15–22 (2019)
29. Tang, W., Yu, P., Wu, Y.: Deeply learned compositional models for human pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 190–206. Springer, Berlin (2018)
30. Lin, T.Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2117–2125. IEEE, Piscataway (2017)
31. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 483–499. Springer, Berlin (2016)
32. Kreiss, S., Bertoni, L., Alahi, A.: Pifpaf: Composite fields for human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11977–11986. IEEE, Piscataway (2019)
33. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988. IEEE, Piscataway (2017)
34. Green, J., Green, P.S.: Alberti's perspective: a mathematical comment. Art Bullet. 69(4), 641–645 (1987)
35. Zhang, Y., Zhou, D., Chen, S., Gao, S., Ma, Y.: Single-image crowd counting via multi-column convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 589–597. IEEE, Piscataway (2016)
36. Drywień, M., Górnicki, K., Górnicka, M.: Application of artificial neural network to somatotype determination. Appl. Sci. 11(4), 1365 (2021)
37. 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 (CVPR), pp. 3686–3693. IEEE, Piscataway (2014)
38. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al.: Microsoft coco: Common objects in context. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 740–755. Springer, Berlin (2014)
39. Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H.S., Lu, C.: Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10863–10872. IEEE, Piscataway (2019)
40. Ross, C.F., Kirk, E.C.: Evolution of eye size and shape in primates. J. Hum. Evol. 52(3), 294–313 (2007)
41. Verkicharla, P.K., Mathur, A., Mallen, E.A., Pope, J.M., Atchison, D.A.: Eye shape and retinal shape, and their relation to peripheral refraction. Ophthal. Physiolog. Optics 32(3), 184–199 (2012)
42. Norton, K., Whittingham, N., Carter, L., Kerr, D., Gore, C., Marfell Jones, M.: Measurement techniques in anthropometry. Anthropometrica 1, 25–75 (1996)
43. Newell, A., Huang, Z., Deng, J.: Associative embedding: End-to-end learning for joint detection and grouping. In: Proceedings of the International Conference on Neural Information Processing Systems (NIPS), pp. 2277–2287. MIT Press, Cambridge (2017)
44. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv preprint, arXiv:190407850 (2019)
45. Sun, K., Geng, Z., Meng, D., Xiao, B., Liu, D., Zhang, Z., et al.: Bottom-up human pose estimation by ranking heatmap-guided adaptive keypoint estimates. arXiv preprint, arXiv:200615480 (2020)
46. Luo, Z., Wang, Z., Huang, Y., Wang, L., Tan, T., Zhou, E.: Rethinking the heatmap regression for bottom-up human pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13264–13273. IEEE, Piscataway (2021)
47. Jin, S., Liu, W., Xie, E., Wang, W., Qian, C., Ouyang, W., et al.: Differentiable hierarchical graph grouping for multi-person pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 718–734. Springer, Berlin (2020)
48. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2961–2969. IEEE, Piscataway (2017)
49. Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: Rmpe: Regional multi-person pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2334–2343. IEEE, Piscataway (2017)
50. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint, arXiv:14126980 (2014)
51. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems (NIPS), vol. 32, pp. 8026–8037. MIT Press, Cambridge (2019)
52. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1385–1392. IEEE, Piscataway (2011)

How to cite this article: Du, C., Yu, H., Yu, L. A scale-sensitive heatmap representation for multi-person pose estimation. IET Image Process. 2021;1–14. https://doi.org/10.1049/ipr2.12404