This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3197364

EHPE: Skeleton Cues-based Gaussian Coordinate Encoding for Efficient Human Pose Estimation

Hai Liu, *Senior Member, IEEE*, Tingting Liu, *Member, IEEE*, Yu Chen, Zhaoli Zhang, *Member, IEEE*, You-Fu Li, *Fellow, IEEE*

***Abstract*—Human pose estimation (HPE) has many wide applications such as multimedia processing, behavior understanding and human-computer interaction. Most previous studies have encountered many constraints, such as restricted scenarios and RGB inputs. To mitigate constraints to estimating the human poses in general scenarios, we present an efficient human pose estimation model (i.e., EHPE) with joint direction cues and Gaussian coordinate encoding. Specifically, we propose an anisotropic Gaussian coordinate coding method to describe the skeleton direction cues among adjacent keypoints. To the best of our knowledge, this is the first time that the skeleton direction cues is introduced to the heatmap encoding in HPE task. Then, a multi-loss function is proposed to constrain the output to prevent the overfitting. The *Kullback-Leibler* divergence is introduced to measure the predication label and its ground truth one. The performance of EHPE is evaluated on two HPE datasets: MS COCO and MPII. Experimental results demonstrate that EHPE can obtain robust results, and it significantly outperforms existing state-of-the-art HPE methods. Lastly, we extend the experiments on infrared images captured by our research group. The experiments achieved the impressive results regardless of insufficient color and texture information.**

***Index Terms*—Human pose estimation, regularization, Gaussian coordinate encoding, skeleton direction, deep learning.**

Manuscript received January 1, 2021; revised May 11, 2022 and July 20, 2022; accepted July 29, 2022. This work was supported This work was supported in part by the National Key R&D Program of China under Grant 2021YFC3340802, in part by the National Natural Science Foundation of China under Grant 62177018, Grant 62173286, Grant 62011530436, Grant 62077020, Grant 62005092, and Grant 61875068, and in part by the Fundamental Research Funds for the Central Universities under Grant CCNU20ZT017 and Grant CCNU2020ZN008. (*Corresponding author: Tingting Liu and Yu Chen*)

Hai Liu, Yu Chen and Zhaoli Zhang are with the National Engineering Research Center for E-Learning, Central China Normal University, Wuhan 430079, China (e-mail: hailiu0204@ccnu.edu.cn; 1585707543@qq.com; zl.zhang@ccnu.edu.cn).

Tingting Liu is with School of Education, Hubei University, No. 368 Youyi Road, Wuhan 430062, Hubei, China, and also with Department of Mechanical Engineering, City University of Hong Kong, Kowloon, Hong Kong. (e-mail: tliu@hubu.edu.cn)

Youfu Li is with the Department of Mechanical Engineering, City University of Hong Kong, Kowloon, Hong Kong, and also with the City University of Hong Kong Shenzhen Research Institute, Shenzhen 518057, China. (e-mail: meyfli@cityu.edu.hk).

I. Introduction

HUMAN pose estimation (HPE) is a long-standing and challenging problem in the field of computer vision, and it aims to locate the keypoints of each person in an RGB image [1-3]. As an upstream task of behavior monitoring [4], human-computer interaction [5-7], action recognition [8, 9], and online learning [10], HPE must be improved in terms of accuracy and robustness [11, 12]. However, HPE often encounters some challenges in real scenarios, thus affecting the accuracy of location. For example, limbs and its proneness to the ambiguity of human body and its background objects are often occluded. To this end, the accurate extraction of human features from images is crucial to the development of the HPE task.

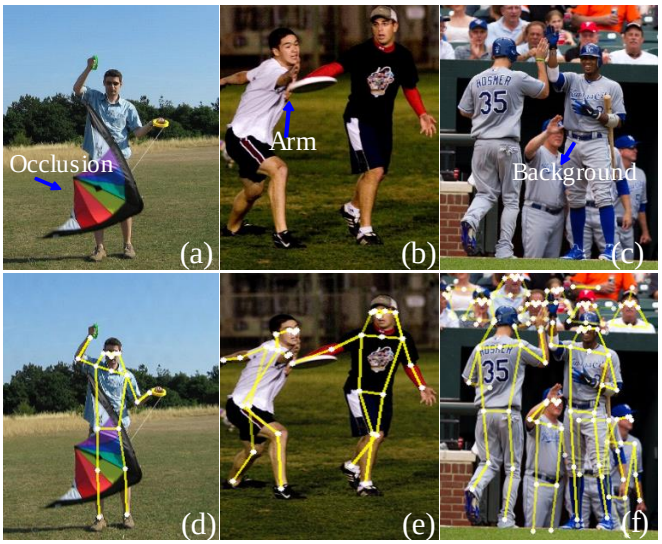

Fig. 1. Human pose estimation results in natural scenes by the proposed EHPE method. (a) A person occluded by a kite. (b) Adjacent interference. (c) Chaotic background. (d)-(f) HPE results of our EHPE method.

The extraction of human pose features can be categorized into handcraft feature-based method (HCFB) and convolutional neural networks-based method (CNNB). In the HCFB method [13], the scale-invariant feature transform and histogram of oriented gradient of the human body parts in the image are calculated first. Then, the graph model is established with the human body joints as nodes. Last but not the least, the state space is continuously reduced by combining the prior constraints of human kinematics. The HCFB method is often

efficient. However, HCFB cannot describe and express the deformation that often occurs in the process of human movement due to the characteristics of hinge and individual differences. Moreover, the generalization ability of these models in natural scenes is unsatisfactory. To solve this problem, a CNN-based method was developed with the rapid development of deep learning technology. Given its excellent performance in image data processing, CNNs have achieved great success when they were introduced into the field of HPE. An important prerequisite is the acquisition of training data for the normal function of the neural network model. The general detection process for the CNN-based HPE model can be divided into three stages, namely, preprocessing of the training data, feeding into the designed network architecture, and post-processing of the output data.

Over the past decades, numerous studies have proposed CNN techniques, such as the cascaded pyramid model [14]. The purpose is to alleviate the problems of keypoints that are difficult to detect. In [15], the predicted heatmap is spatially accurate by connecting high- and low-resolution subnets in parallel. In our previous work [16], a light multi-stream neural network is proposed to learn the view-invariant representations from skeletal self-similarities of varying scales for the human action recognition. To address the detection problem occluded keypoints, Zhang *et al.* [17] designed an efficient network structure named cascaded context mixer, with three useful training strategies and four effective post-processing techniques. These methods have made outstanding contributions to the learning of multi-scale features of images.

Although CNNs are powerful in learning image features and have been successfully leveraged to numerous of computer vision tasks [18], three fundamental issues for CNN-based HPE remain.

1) Object occlusion: The occlusion of the predicted human body frequently occurs in real scenarios (see in Fig. 1(a)). Occlusion by objects results in the loss of partial joint information, which affects the acquisition of local features by the network model.
2) Neighbor interference: Generating ambiguity in CNN-based HPE models is easy when classifying keypoints due to the similar texture, color and structure features among human bodies. In Fig. 1(b), the interference between adjacent human joints plays a vital role in the blocking localization effect.
3) Complex background: Human body detection is disturbed by a complex background in general cases (see in Fig. 1(c)), making localization difficult. Images dominated by human bodies mixed with background can be hardly recognized by the CNN, regardless of the strong contrast between the foreground and background.

The above three challenges can be summarized as a problem of location interference, including interference with objects, adjacent characters, and background and the loss of important information. These problems make the development of the HPE task limited to a simple scenario without these interference factors. Hence, there is an urgent need to develop HPE methods that can perform well in general scenarios.

Natural scenes have a strong spatial geometric relationship between the adjacent keypoints of the same individual and thus can provide a strong basis for inferring the position of the keypoints of the human body that are disturbed by the above. Hence, in addition to constructing a network model for implicit learning, contextual information and high-level semantic relationships of the human pose must be investigated to improve the accuracy of HPE further. However, these topics are ignored by the method proposed in [14], [15], [17]. Furthermore, to the best of our knowledge, no research has introduced this keypoint object-level relationship into the HPE task. Therefore, the motivation of our work is to design an encoding method that can adapt to the change of the limb direction and the position of adjacent keypoints. Meanwhile, it can combine the ability of the network to judge remote spatial context information. In summary, the location of human keypoints is predicted by considering the important effects of local keypoints and high-level semantic relationships. To achieve this goal, we innovatively construct a new human body keypoint encoding method that can automatically explore body direction clues. Compared with that of previous works, the major contributions of this study can be summarized in three aspects.

1) A novel HPE model is developed to reveal effectively the skeleton direction cues for keypoint coordinate encoding. The anisotropic Gaussian label is constructed for each keypoint in accordance with adjacent limb connection. To the best of our knowledge, this is the first study to introduce skeleton direction information in the heatmap encoding of HPE task.

2) An efficient yet robust convolution neural network architecture is proposed optimized by the KL and L_2 norm loss (multi-loss). This novel loss can effectively measure the difference between ground-truth heatmap distribution and predicted one.

3) Experiments are conducted on two datasets: MS COCO and MPII datasets, and the extended experiment is carried out on the infrared images captured by our group. Compared with several baseline and state-of-the-art methods, the proposed EHPE shows better performance, thus validating its effectiveness.

The rest of this article is organized as follows. The related work of this article is described in Section II. In Section III, the proposed method is introduced in detail. Furthermore, the experiments and results are illustrated in Section IV, we conclude this research in Section V.

II. Related Work

A. CNN-based HPE Model

The HPE task aims to predict the keypoints of one or more persons in an image/video and thus is beneficial for understanding human action and intentions. In the era of artificial intelligence, an HPE model based on deep learning technique has emerged. In contrast to the handcraft-based method, CNN-based architecture can obtain the global context information of images and capture the multiscale joint point feature vectors in different receiving fields. Therefore, it can extract the scene information closest to the real one. Recently, many HPE methods have been proposed. Toshev *et al*. [19] first

proposed an HPE formula based on a deep neural network. Tompso *et al*. [20] proposed a hybrid architecture including the Markova random field and CNN. Chu *et al.* [21] leveraged the hourglass as the baseline and introduced the conditional random field instead of global features for the spatial correlation modeling. Based on the improvement of the stacked hourglass network, Ke *et al*. [22] proposed a multi-scale regression network (MSR-net) and a multi-scale monitoring network (MSS-net). It combines rich multiscale features and improves the robustness of keypoint location through cross-scale feature matching in comparison with [19, 21, 23]. To solve and study human frame inaccuracy and pose estimation in crowded scenes, several approaches have been proposed, such as the regional multi-person pose estimation (RMPE) [24] method, and the CrowdPose [25] method. Nie *et al*. [26] proposed a novel analytically induced learner, which assists HPE by effectively applying analytical information of body parts. Some studies were also devoted to lightweight networks.

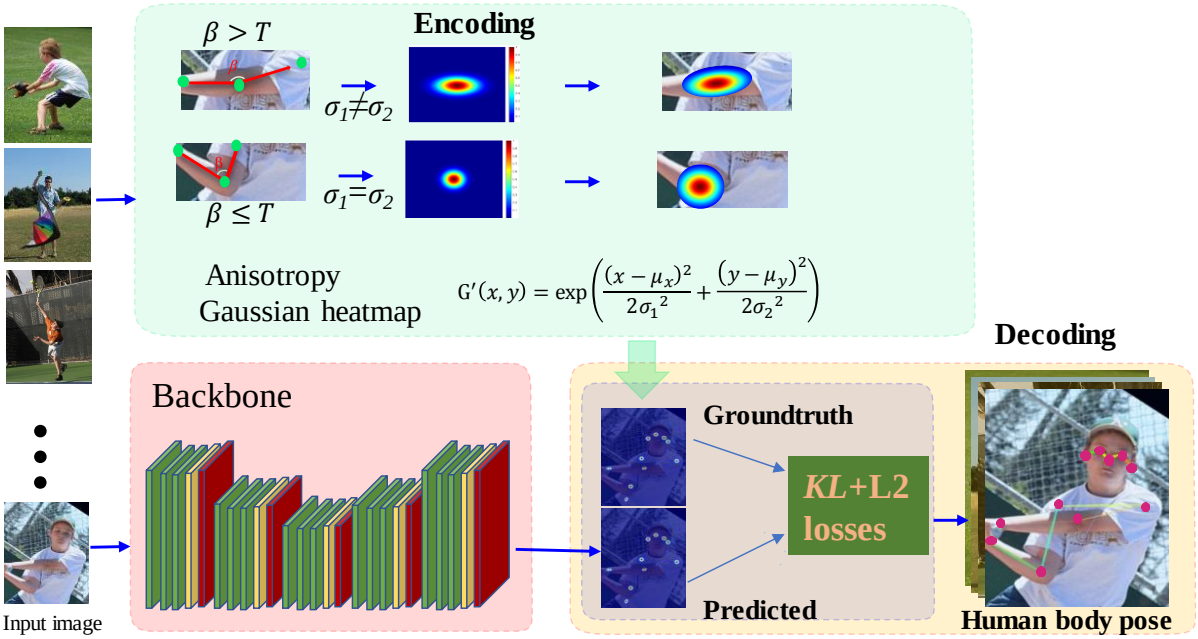

Fig. 2. Concrete architecture of the proposed EHPE neural network. The symbol β is angle of different limbs (lower limb and upper). T is the threshold for performing anisotropy coding.

For example, the simple baseline model [27] proposed by Xiao *et al.* is more intuitive and simpler than the stacked hourglass model [28]. The main contribution of Fastpose [29] proposed by Zhang *et al.* is its light weight network that applies knowledge extraction to the detection of human body keypoints. In [15], Wang *et al.* constructed a high-resolution network, which can interfuse multi-scale features in the entire network and achieves the optimal performance. Artacho *et al.* [30] proposed a void space pooling module based on the waterfall model, which is a unified framework independent of post-processing. Many other kinds of CNN-based methods still attract research in this community. Although existing studies have made some progress in HPE, no work has considered learning the relationship among adjacent limbs in a person to predict body poses.

B. Heatmap Regression

The encoding of HPE method can be divided into two kinds of channels. The first is regression based on labels (direct regression), and the other is regression based on heatmap (heatmap-based). For label-based regression, Toshev *et al.* proposed the cascading DNN regression [19] to predict human keypoint in a holistic way. Carreira *et al.* [31] used this approach of direct regression in their respective studies. Liu *et al*. [32] proposed a nonuniform Gaussian-label distribution learning method for the head pose estimation task, which aims at predicting the orientations of head pose. One advantage of adopting this framework is that it allows for end-to-end learning and continuous output. However, without other processes, it is very difficult to learn the mapping characteristics directly from the original image. In other words, the direct regression to the labeled coordinates is unstable and will cause large fluctuations. Therefore, the network hardly converges. However, the regression of heatmap-based methods solves the shortcoming of direct regression because of two reasons. On the one hand, the keypoints of a human body cannot be accurately defined by a certain pixel; thus, the heatmap-based method can overcome the problem of inaccurate data annotation. On the other hand, the human body priori can reveal that adjacent key points have a strong correlation, and this interrelated nature is difficult to capture by regressing coordinates independently. Intuitively, the heatmap-based method converts the target pixel into a probability distribution area and performs classification before regression, thus greatly reducing the difficulty of convergence of the model.

For heatmap-based regression, the convolutional pose machine [33] performs multistage regression on the heatmap. It learns remote connection nodes by amplifying the receptive field. Intermediate supervision is also utilized to avoid gradient disappearance. In [28], the author designed a stacked hourglass model, whose major contribution lies in the use of multiscale features to identify posture. Some works also considered the importance of coordinate representation. Zhang *et al.* [34] used

the maximum value on the heatmap and its corresponding position to estimate the mean position of the true Gaussian distribution, the quantization error caused by lower sampling can be reduced to the greatest extent. In [35], Huang *et al.* designed continuous metrics in their work to eliminate the inherent errors caused by using pixel distance metrics in affine transformations. These works have achieved good results from their motivations, but the inflexible tag construction method has been adopted, and the final effect of its estimation is still limited by some frequent problems of HPE. In view of this situation, it is reasonable to believe that an encoding method that conforms to the human a priori needs to be proposed urgently. In this manner, the network returns the predicted value of the skeleton direction as far as possible when making the predictions, thus helping the network understand the semantic interconnection between the keypoints easily.

III. Architecture of Proposed Model

A. Overview of EHPE Model

The overall model architecture is shown in Fig. 2. The proposed EHPE method includes three modules. The first module is the network model used for training parameters. The second module is the encoding module of the input image, and the third module realizes the decoding of the output image. In this work, the HRNet is selected as the backbone.

However, network architecture must be modified for improved efficiency and performance. The developed network consists of three layers: the convolutional layer, the covariance pooling layer, and the output layer. For the convolutional layer, HRNet [36] is chosen as the backbone to extract the features of the input images. Traditional CNNs are designed with convolutional layers, pooling layers, and FC layers to capture only the first-order statistics, such as the mean or maximum of the eigenvalues. Second-order statistics, such as the covariance, are deemed to be better regional descriptors than first-order statistics [37]. The core of the HPE task is directly bound up with how human keypoints are distorted, rather than the detection of their presence. Obtaining second-order statistics is more suitable than using first-order statistics to present such distortions. Thus, we introduce covariance pooling instead of average or maximum pooling after the last convolutional layer and build covariance matrices as global image representations. Backpropagation is not easy due to the nonlinear functions involved in covariance pooling. Therefore, end-to-end learning [38] is referred to calculate the gradients.

B. Encoding with the Anisotropic Gaussian

For the HPE, coordinate encoding is an indispensable part of the heatmap-based approach. In general, the Gaussian function is used to encode the annotation keypoints in the dataset. In this study, a novel encoding scheme is proposed, and it can be summarized into three stages: classification, fitting, and encoding (CFE). The pipeline of label construction is shown in Fig. 3. The first stage is to construct anisotropic and isotropic Gaussian distributions near the labeled points in accordance with the similarities and differences of variances. The second stage illustrates the process of anisotropic distribution to fit the limb direction. In the third stage, we present the final differentiated encoding strategy. In accordance with the above steps, the details are described below.

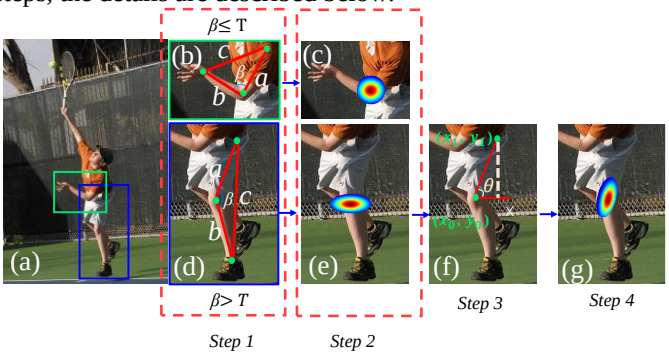

Fig. 3. Pipeline of label construction. (a) Original HPE image. (b)-(c) Gaussian label in green box. (d)-(g) Elliptic Gaussian label in blue box if the angle is larger than the given threshold *T*.

***Stage I**-Classification of heatmaps*: In this part, a new classification of heatmaps is revealed. For a given key point coordinate (μ_x, μ_y), it is used as the center to generate a Gaussian heatmap. For the limbs in Fig. 3(a), the label is constructed as

$$G(x,y) = exp\left(\frac{1}{2}(x-\mu)^T\boldsymbol{\Sigma}^{-1}(y-\mu)\right), \quad (1)$$

$$\mu = \begin{bmatrix}\mu_x\\ \mu_y\end{bmatrix}, \quad (2)$$

where $\sum$ is the covariance matrix, which is utilized to generate multivariate Gaussian heatmap. It is indicated by the following formula,

$$\Sigma = \begin{bmatrix}{\sigma_1}^2 & 0\\ 0 & {\sigma_2}^2\end{bmatrix}. \quad (3)$$

where ${\sigma_1}^2$ and ${\sigma_2}^2$ represent the variance of the Gaussian heatmap. Figure 3(c) shows the Gaussian heatmap when ${\sigma_1}^2 = {\sigma_2}^2$. It appears as an isotropic circle in the image. As shown in Fig. 3(e), we can intuitively observe that the Gaussian function presents an elliptical shape if ${\sigma_1}^2 \neq {\sigma_2}^2$. Specifically, the default setting is ${\sigma_1}^2 > {\sigma_2}^2$.

***Stage II**-Fitting of skeleton direction:* To establish a Gaussian label in line with the priori of human body structure as much as possible, the orientation of the skeleton must be fitted at this stage.

Given the anisotropy of Gaussian distributions whose variances are different, we need to rotate the entire Gaussian probability distribution to follow the direction of the upper limb in the image space. In the process, each point on the distribution is adjusted from (x, y, ω) to (x', y', ω); ω represents the coding probability value at this coordinate keypoint. (x', y') is the position of (x, y) rotated around the key point coordinates. The rotation operation can be given by the following formula,

$$\begin{bmatrix}x'\\ y'\\ 1\end{bmatrix} = M\begin{bmatrix}x\\ y\\ 1\end{bmatrix}, \quad (4)$$

where the matrix M is an affine transformation matrix. When a pixel with a probability value is ready to be rotated in the image space, the origin coordinate must be moved first to the center of the rotation. Then, the rotation matrix is performed by the rotation operation. Lastly, the rotated result is mapped back to the original coordinate space. In accordance with the above three steps, the matrix M can be expressed by the following expression:

$$M = \begin{bmatrix} 1 & 0 & \mu_x \\ 0 & 1 & \mu_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & -\mu_x \\ 0 & 1 & -\mu_y \\ 0 & 0 & 1 \end{bmatrix}, \quad (5)$$

where the calculation of the rotation angle θ is the key step. As mentioned above, the initially constructed anisotropic Gaussian heatmap exhibits an elliptical shape. Its long axis is in the x direction of the image coordinate system. In Fig. 3(f), the keypoint adjacent to the current keypoint must be retrieved. The rotation angle θ can be expressed by the following formula:

$$\theta = \begin{cases} \tan^{-1}\left(\frac{y_0 - y_1}{x_1 - x_0}\right), & \text{if } x_1 \neq x_0; \\ \frac{\pi}{2}. & \text{otherwise.} \end{cases} \quad (6)$$

where (x_0, y_0) is the keypoint of the current encoding, and (x_1, y_1) refers to the coordinate of the upper adjacent point of the current encoding keypoint. The adjacent keypoint of the upper bone of each keypoint is given by the dictionary definition. As shown in Fig. 3(g), we can fit the orientation of the skeleton.

Stage III*-Encoding of anisotropic strategy:* In this stage, the final encoding scheme based on the above method is given. Before that, we provide differentiated treatment for keypoints in different states.

In Figs. 3(b) and 3(d), the law of cosines is introduced to calculate the angle formed between the current keypoint and the adjacent limb. Based on the currently encoded keypoint, a and b represent the length of the upper and lower limbs, and c describes the straight-line distance between the upper and lower adjacent keypoints. The method of calculating the angle between the limbs is given by

$$\beta = cos^{-1}\left(\frac{a^2 + b^2 - c^2}{2ab}\right) \quad (7)$$

In *Step*_1 in Fig. 3, the included angle β provides the premise for implementing the anisotropic encoding strategy. Then, the overall label generation scheme is illustrated as,

$$EG(x,y) = \begin{cases} exp\left(\frac{1}{2}(x' - \mu_x)^T \boldsymbol{\Sigma}^{-1}(y' - \mu_y)\right), if\ \beta > T, \\ exp\left(\frac{1}{2}(x - \mu_x)^T \boldsymbol{\Sigma}^{-1}(y - \mu_y)\right), otherwise. \end{cases} \quad (8)$$

where T is the threshold for performing anisotropic coding. In our experiment, T achieves the optimal performance when it is set at 130 °. The covariance matrix $\sum$ varies in accordance with different alternatives.

C. Decoding

The coordinate decoding method includes two steps. The first step is to obtain the maximum response position m and the second maximum response position s of the output heatmap, Then, d times of the unit length are shifted from the maximum activation position to the second largest activation position. The predicted position p is calculated by

$$p = m + d\frac{s - m}{||s - m||_2}, \quad (10)$$

where the offset distance d is to compensate for the quantization error, with its value usually set as 0.25, which is determined by the expected error. Given that the heatmaps are calculated in the low-resolution pixel space, mapping the coordinates back to the original image space through upsampling is necessary. The final prediction point can be expressed as,

$$\hat{p} = \lambda p, \quad (11)$$

where λ represents the upsampling rate. However, this decoding method based on statistical error cannot realize the practical consideration of the output heat map, thus, we adopt the decoding method based on Taylor distribution perception in this study. To eliminate the influence of multiple peaks, the Gaussian kernel is adopted to smooth the predicted heatmap because the predicted heatmap usually does not show the Gaussian properties well. Then, the keypoints are predicted in accordance with the actual distribution information. Lastly, we restore the predicted position to the original image space and obtain the final results by (11). Interested readers can refer to [34] for decoding in detail.

D. MAP-based HPE Model

Recently, the maximum a posteriori (MAP) estimation method has been widely used for image regression tasks; it uses a *prior* probability density functions as the prior constraints. It has played a key role in tracking the ill-posed problems widely existing in HPE tasks. In this study, the MAP framework is introduced to address the aforementioned problem in HPE for the first time. Given a group of pose images X with their ground-truth heatmap distribution P, the aim of training is to find the best θ estimation by maximizing the posterior probability $p(\theta|X, P)$. The neural network parameters are represented by the symbol θ, which is needed for calculation in EHPE network. For the i-th person instance, the k-th joint point is encoded into the ground truth heatmap $G_k^i(x, y)$. The MAP estimation can be illustrated as,

$$\theta^* = \arg max\ p(\theta|X, P). \quad (12)$$

Based on Bayes criterion, (12) becomes,

$$\theta^* = \arg max \frac{p(X,P|\theta)p(\theta)}{p(X,P|\theta)}. \quad (13)$$

Since $p(X, P)$ is independent of the variable θ, and $p(X, P)$ can be considered a constant. Thus, (13) can be rewritten as,

$$\theta^* = \arg max\ p(X, P|\theta)p(\theta). \quad (14)$$

The monotonic logarithm function can be rewritten as

$$\theta^* = \arg max\ \log p(X, P|\theta) + \log p(\theta). \quad (15)$$

Two probability density functions need to be defined. The likelihood probability $p(X, P|\theta)$ represents the distance between the predicted distribution and the ground truth distribution. Kullback-Leibler (*KL*) divergence is selected to measure the distance. Consequently, the likelihood probability can be presented as,

$$p(X, P|\theta) = \sum_j p_j ln\frac{p_j}{g_j}, \quad (16)$$

where g denotes the prediction heatmap value in G. Added with the *KL* divergence, the loss function is proposed as,

$$L(\theta) = \frac{1}{K}\sum_k^K \sum_i^I (P_k^i\ ln\frac{P_k^i}{G_k^i}) + \eta\|\theta\|^2, \quad (17)$$

where P_k^i, G_k^i represent the predicted value and the true value on the heatmap, respectively.

To reduce the overfitting issue, the Euclidean distance is introduced to measure the predicted heatmap distribution P_k^i and ground-truth heatmap distribution G_k^i. Then, the proposed loss is defined as,

$$L(\theta) = \frac{1}{K}\sum_k^K \sum_i^I (P_k^i\ ln\frac{P_k^i}{G_k^i}) + \frac{\lambda}{2}\left\|P_k^i - G_k^i\right\|^2 + \eta\|\theta\|^2. \quad (18)$$

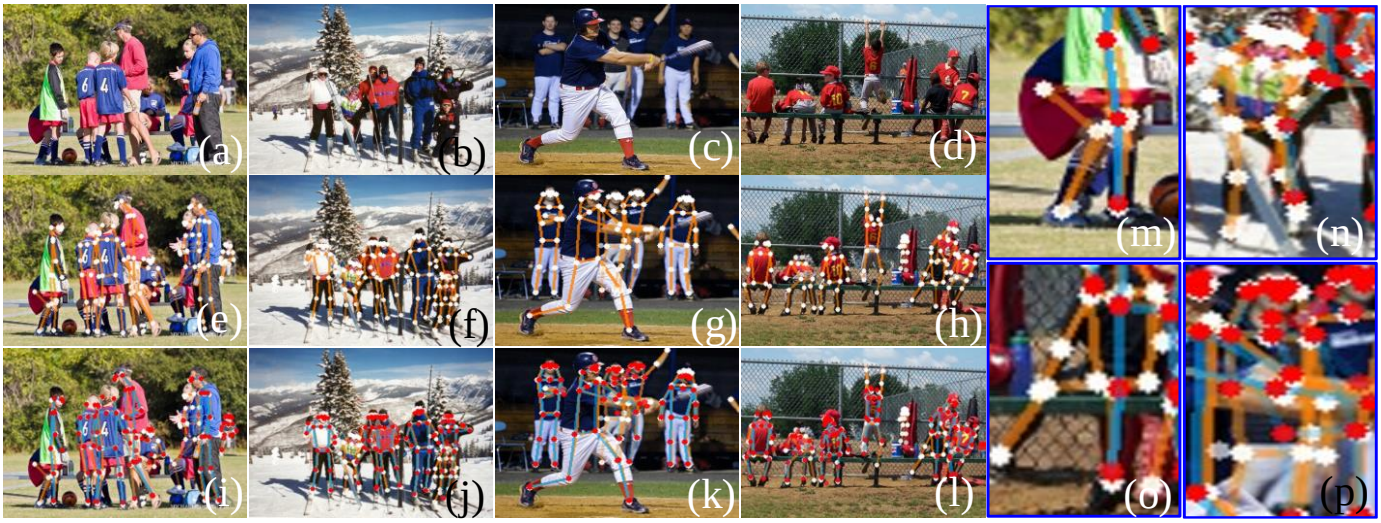

Fig. 4. Comparison of the results. (a)-(d) Original images. (e)-(h) Results by our EHPE method. (i)-(l) Contrast between our method (orange line) and DARK (blue line), with the details shown in the blue boxes (m)-(p).

The derivative of $L(\theta)$ with respect to parameters θ can be rewritten as,

$$\frac{\delta L(\theta)}{\delta p_k} = g_k + (\lambda p_k - \lambda g_k - 1)\frac{\exp(p_k)}{\sum_k \exp(p_k)}. \quad (19)$$

This updated formulation is easy to vectorize for training batch input. The Adam method is introduced to minimize the objective loss function $L(\theta)$, and its optimization process is provided in **Algorithm** 1.

Algorithm 1. Training strategy for the proposed EHPE model.

Input: Human pose image $\boldsymbol{X}$ in the training set.
Set: Batch size t, learning rate α, exponential decay rate of the first moment estimation φ_1, exponential decay rate of the second moment estimation φ_2, parameter ε is set as a small positive constant.
i) Initialize parameter vector θ_t,
Initialize biased first moment estimation m_0,
Initialize biased first moment estimation v_0;
ii) **While** θ_t not converged **do:**
$t \leftarrow t+1$, $m_0 \leftarrow 0$;
Compute gradients w.r.t. loss function g_t
$m_t \leftarrow \varphi_1 \cdot m_{t-1} + (1-\varphi_1) \cdot g_t$;
$v_t \leftarrow \varphi_2 \cdot m_{t-1} + (1-\varphi_2) \cdot g^2_t$;
$\hat{m}_t \leftarrow m_{t-1} / (1-\varphi^t_1)$;
$\hat{v}_t \leftarrow v_t / (1-\varphi^t_2)$;
$\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \hat{m}_t / (\sqrt{\hat{v}_t} + \varepsilon)$;
end while
Output: Optimization parameters θ_t

IV. Experimental Results and Discussions

A. General Setting

1) Datasets: Two of the most widely used public datasets are introduced in our experiments, including the MSCOCO dataset developed by Microsoft and the MPII dataset with both single-player and multiplayer scenarios.

MS COCO [39]: This database is collected for the common objects in the complex field scenario by Microsoft. It is a large, rich object detection, segmentation and subtitle dataset. This famous dataset is used widely for many tasks, such as segmentation, keypoint detection, object detection, and subtitles. In this study, a dataset of keypoint detection is adopted in our experiments. In the MS COCO2017 dataset, the training set contains 118,287 images, whereas the test set contains 5,000 images. Each instance has 17 keypoints, including the nose and eyes. Annotations on train and validation (with over 1.7 million labeled keypoints on 150,000 subjects) are publicly available. In evaluation, we followed the commonly used train2017/val2017/test-dev2017 split.

MPII *dataset* [40]: It contains about 25K images, a total of 40K human body instances, and 16 keypoints for each image. The annotations of training and validation sets are publicly benchmarked. The images in the MPII data set are extracted from YouTube videos and can be used as HPE for single person and multiple persons. In our experiment, the standard train/val/test split is adopted as in [20].

2) Baseline methods: To validate the developed EHPE approach, several state-of-the-art methods are selected in the comparison experiments.

- **RMPE** [24]: A symmetric spatial transformation network is proposed to strengthen human object instances.
- **Mask-RCNN** [41]: One-hot coding is performed on K keypoints of the human body, and the types of K masks are predicted to achieve pixel-level segmentation.
- **OpenPose** [42]: Part affinity fields are introduced into the bottom-up HPE task.
- **CPN** [14]: The cascade pyramid network can adopt appropriate methods for the key points with different recognition difficulty levels.
- **CFN** [43]: The network leverages multi-level supervision to realize the keypoint location function.
- **Simple Baseline** [27]: A baseline that is simple and reaches SOTA level is proposed.
- **HRNet** [15]: The network uses a multi-scale fusion method to maintain high resolution characterization throughout the entire process.
- **DARK** [34]: An efficient coordinate decoding based on Taylor expansion is proposed.
- **UDP** [35]: Quantitative analysis of system errors introduced by biased data processing, and proposed an unbiased data processing flow.

3) Evaluation metrics: The HPE model has two evaluation metrics, namely, percentage of correct keypoints (*PCK*) and object keypoint similarity (*OKS*).

OKS is used as an evaluation index for MS COCO human body keypoint detection. It aims to calculate the truth value and predict the similarities of human body keypoints. For a human body instance p, the keypoints are written as,

$$OKS_p = \frac{\sum_i exp\left\{-d_{pi}^2/2S_p^2\sigma_i^2\right\}\delta(v_{pi}=1)}{\sum_i \delta(v_{pi}=1)}, \quad (20)$$

where d_{pi} represents the Euclidean distance between the predicted position and the ground truth keypoint, s_p indicates the scale of the human object, v_{pi} indicates the visibility of the keypoint label, and σ_i is the offset of the artificially labeled position. Given OKS threshold s, the average accuracy rate (AP) can be calculated as,

$$AP^s = \frac{\sum_p \delta(OKS_p > s)}{\sum_p 1} \quad (21)$$

PCK [40] represents the proportion of correct keypoints estimated. It is used as the MPII evaluation index to calculate the ratio of the normalized distance between the groundtruth keypoints and its predicted ones less than the given threshold. In the MPII dataset, the head length is utilized as the normalized reference. Thus, it is also called PCKh.

B. Implementation Details

Our experimental environment is on a personal computer server equipped with an NVIDIA TITAN RTX-24G GPU and 64 GB Intel(R) Core(TM) I9-9900K CPU @3.60ghz. Our model implements Pytorch as a framework for deep learning and is trained with 200 epochs. Adam [44][45] is used as optimizer, and the batch size is set as 144. During the experiment, we set the learning rate to 0.001 and attenuate it to one-tenth of the original in 170–200 epochs. In addition, we set the exponential decay rates φ_1 and φ_2 of the moment estimation as 0.99 and 0 respectively in the optimization function.

C. Experiment Results and Analysis

1) Results on the MS COCO Database

The visualized HPE results shown in Fig. 4 verify our success and shows the impressive robustness of our approach. Figs. 4(a)–(d) are the original images, and Figs. 4(e)–(h) show the effect of positioning by our method (orange lines). In Figs. 4(i)–(l), the cyan line is used to show the effect of the comparative method, whereas the orange line shows our predictions. As shown in the details in the blue boxes, our method is surprisingly robust. DARK and our models are also based on CNNs while following the heatmap regression. However, DARK tends to fail to locate the keypoints correctly. The correlation mapping of keypoints generated by the DARK are not prominent, implying that the DARK model is weak in learning and combining high-level contextual keypoint information complements. Quantitative results on the MS COCO dataset are shown in Table I. The final AP accuracy reaches 79.1% and 3.3% higher than that of the state-of-art UDP method. Compared with the previous methods, our model achieves huge improvement, suggesting that our model is successful in exploiting and combining the complements residing in the skeleton direction.

Figure 5 shows the visual results of the proposed label construction, indicating that our tags can adapt to changes of the skeletal direction. Furthermore, we visualized the heatmap obtained using our unique label construction method. In Fig. 5, we show a group of output images during the experiment. According to the proposed rules, the athletes keypoints in Fig. 5(b) are encoded in two ways: the shoulders pointed by the green arrow adopt an isotropic Gaussian coordinate encoding, whereas the knees pointed by the red arrow adaptively use an anisotropic multivariate Gaussian coordinate encoding.

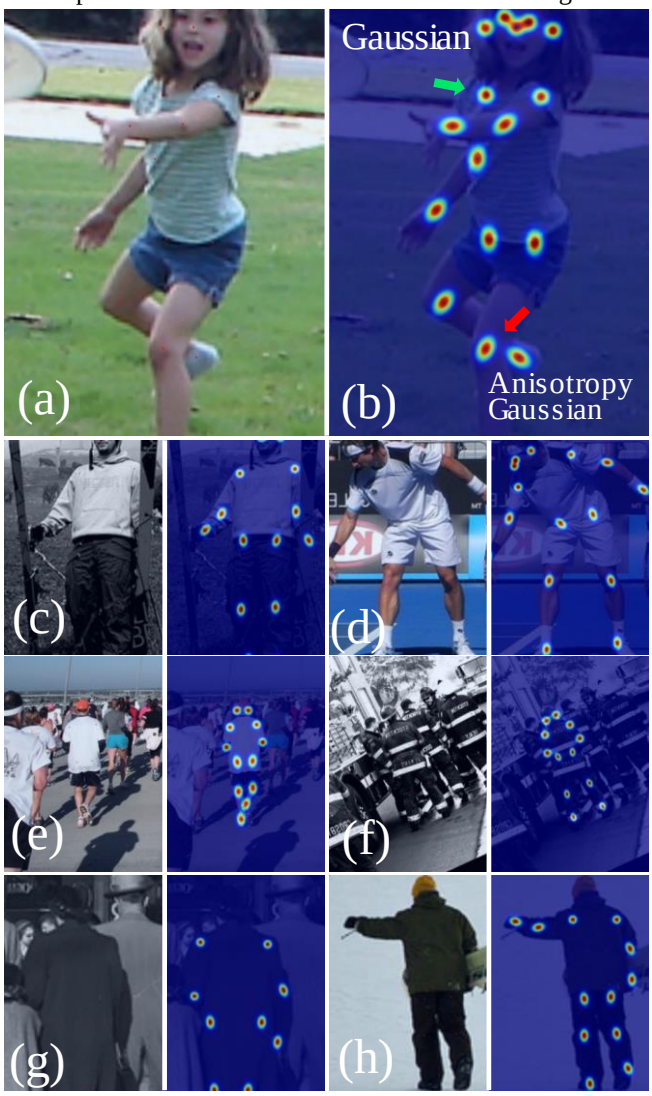

Fig. 5. Visualized results of our label construction by the proposed method on MS COCO. (a) Original HPE image, the girl plays frisbee. (b) anisotropic Gaussian labels.

TABLE I. Comparison result with the state-of-the-art HPE methods on the COCO test-dev set.

Methods	AP	AP^{50}	AP^{75}	AP^{m}	AP^{l}
RMPE [24]	61.8	83.7	69.8	58.6	67.6
Mask-RCNN [41]	63.1	87.3	68.7	57.8	71.4
OpenPose [42]	65.3	85.2	71.3	62.2	70.7
CPN [14]	72.1	91.4	80.0	68.7	77.2
CFN [43]	72.6	86.1	69.7	78.3	64.1
Simple Baselines [27]	73.7	91.9	81.1	70.3	80.0
HRNet [15]	75.3	89.3	82.6	71.4	80.5
DARK [34]	75.5	91.0	82.6	71.5	81.0
UDP [35]	75.8	91.2	83.3	72.1	81.3
EHPE (Ours)	**79.1**	**93.6**	**85.8**	**76.3**	**84.0**

2) Results on the MPII Database

The proposed method was compared with HRNet [15], DARK [34] and UDP [35] on the MPII verification dataset. As shown in Table II, even on the more stringent PCK 0.1 measure, our method showed excellent performance. The number of samples provided by MPII is far less than that of the MS COCO dataset, indicating that our method can train datasets with multi-resolutions.

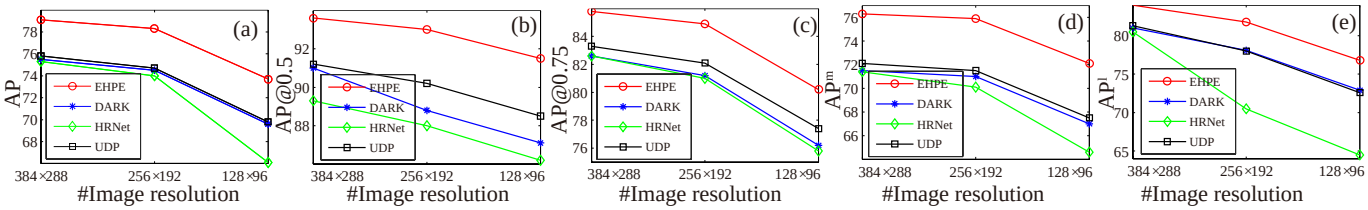

Fig. 6. Comparison of the detection accuracy with the image resolution decreasing for (a)AP. (b) AP@0.5 (c)AP@0.75 (d)AP^m (e)AP^l

D. *Ablation Study and Discussion*

In this section, we discuss the effectiveness of each module of this method in depth. Our experimental design is described as follows. First, we conducted ablation experiments to prove the individual effectiveness of each module. Then, we compare the state-of-the-art methods in input images of different sizes. Next, our model is placed on different backbones for experiments to demonstrate that our method can be seamlessly integrated into any network. Lastly, we discussed the effect of the threshold of the keypoint label generation scheme and the coordinate encoding variance ratio on the final result.

TABLE II. PCKh evaluation by the HRNet, DARK, UDP, EHPE methods on the MPII dataset.

Methods	Head	Kne.	Hip	Wri.	Elb.	Sho.	Ank.	PCKh
PCKh@0.5								
HRNet	97.1	**87.1**	89.1	86.5	90.3	95.5	83.3	90.3
DARK	97.2	86.7	89.7	86.7	91.2	95.9	84.0	90.6
UDP	**97.4**	86.5	89.1	86.5	90.8	**96.1**	83.3	90.4
EHPE	**97.4**	86.8	**90.0**	**86.8**	**91.4**	96.0	**84.2**	**90.8**
PCKh@0.1								
HRNet	51.1	29.9	17.9	41.6	42.0	42.7	31.0	37.7
DARK	55.2	**33.4**	20.1	45.2	47.4	47.8	35.4	42.0
UDP	55.3	33.3	20.2	45.4	47.5	47.9	35.6	42.1
EHPE	**55.5**	33.2	**20.3**	**46.1**	**47.8**	**48.2**	**36.3**	**42.3**

1) Module

In Table III, we conducted ablation experiments on each module in EHPE. In this set of experiments, the size of the input image is fixed to 384 × 288, and then the CFE module is removed and the multi-loss is replaced with MSE. It is not difficult to see that the accuracy of the model is gradually decreasing. The model reached the highest accuracy when it was fully equipped with the method we proposed. We first demonstrate the importance of the skeleton direction cues-aware in the location interference problem. For comparison, the CFE module is removed, and the quantitative results compared between row 1 and 2 in Table III show the considerable boost by involving the CFE module into EHPE. Then, the multiloss is replaced with MSE. Meanwhile, we eliminate the regularization learning strategy. The accuracy of the model gradually decreases to a certain extent. The reasons should be attributed to the strategy that uses *KL* divergence as supervised learning method and the adopted regularization learning. Training with the main loss of *KL* divergence instead of a single loss can make the output more closely approximate the label we constructed. Furthermore, the strategy of regularization learning prevents the network from overfitting and converging to a lower accuracy. λ is obtained in a data-driven manner. Lastly, except for the modification of the backbone network, all other modules have been removed. Nevertheless, our method is still working at their best. The effectiveness of each module in EHPE is verified. In the revised architecture, we remove or add the module in EHPE, and the corresponding variants are named as "w/o module" and "w/ module", respectively.

TABLE III. Results of ablation experiments on each module.

Methods	Input size	AP	AP^{50}	AP^{75}	AP^m	AP^l
w/o CFE+multi-loss	384×288	75.5	91.0	82.6	71.5	81.0
w/o multi-loss	384×288	76.1	91.3	82.9	72.0	81.8
w/o CFE	384×288	77.2	92.1	83.5	73.6	82.1
EHPE	384×288	**79.1**	**93.6**	**85.8**	**76.3**	**84.0**

2) Effect of the Image Resolution

In vision community, image resolution is a major factor affecting the final prediction results. Under normal circumstances, the accuracy of keypoint positioning shows a downward trend as the image resolution decreases. Therefore, the output results of different input image sizes are discussed. Three different resolutions of images are used as model inputs, including 128×96, 256×192, and 384×288. Table IV shows the comparison results of our model and state-of-art method under different resolution conditions. The results show the robustness of our proposed method under the influence of image resolution. The EHPE method outperformed the existing models from high-resolution to low-resolution.

Moreover, we also discuss the effect of model accuracy reduction with the decreasing of the image resolution. As shown in Fig. 6, when the image size is reduced from 384×288 to 256×192, the AP is only reduced by 0.8%, which is more stable than UDP, DARK and HRNet; when the image resolution continues to decrease to 128×96, the AP of EHPE decreases by 4.6 %, which is lower than the AP accuracy reduction of UDP (4.9%), DARK (5.0%) and HRNet (7.9%), respectively. In Table IV, experiments have proven that our model has less performance loss on low-resolution images, thus providing support for deploying human pose estimators on low-resource devices.

TABLE IV. Comparison results by several state-of-the-art models with the image resolution increasing.

Methods	Resolution	AP	AP^{50}	AP^{75}	AP^m	AP^l
HRNet-W32 [15]	128×96	66.1	86.2	77.4	64.6	64.5
DARK[34]	128×96	69.6	87.1	76.2	67.0	72.9
UDP[35]	128×96	69.8	88.5	78.2	67.5	72.6
EHPE (ours)	128×96	**73.7**	**91.5**	**80.2**	**72.1**	**76.8**
HRNet-W32[15]	256×192	74.0	88.0	81.2	70.1	70.5
DARK[34]	256×192	74.5	88.8	81.2	71.0	78.1
UDP[35]	256×192	74.7	90.2	82.1	71.5	78.0
EHPE(ours)	256×192	**78.3**	**93.0**	**84.9**	**75.9**	**81.8**
HRNet-W32 [15]	384×288	75.3	89.3	82.6	71.4	80.5
DARK[34]	384×288	75.5	91.0	82.6	71.5	81.0
UDP[35]	384×288	75.8	91.2	83.3	72.1	81.3
EHPE (ours)	384×288	**79.1**	**93.6**	**85.8**	**76.3**	**84.0**

3) Effect of Backbone

To verify the network structure independence of the model, experiments were conducted on different CNN architectures. As shown in Table V, we combined the model with HRNet [15], ResNet, and Hourglass, respectively. We conducted it with different evaluation dimensions on different input sizes. The obtained accuracy is higher than the existing excellent methods. It demonstrates that the proposed EHPE method can be seamlessly integrated into any existing backbones.

TABLE V. Comparison results with different backbone networks, such as HRNet and ResNet. The underlining indicates the second-best effect. "cp" denotes the covariance pooling. "Taylor" is the Taylor distribution perception in the decoding stage.

Methods	Resolution	AP	AP^{50}	AP^{75}	AP^{m}	AP^{l}
Hourglass	128×96	66.2	87.6	75.1	63.8	71.4
Hourglass+DARK	128×96	69.6	87.8	77.0	67.0	**75.4**
Hourglass+EHPE	128×96	**70.1**	**88.2**	**77.2**	**69.3**	<u>75.2</u>
ResNet	128×96	59.3	85.5	67.4	57.8	63.8
ResNet+DARK	128×96	62.6	86.1	70.4	60.4	**67.9**
ResNet + EHPE	128×96	**62.8**	**88.3**	**70.5**	**61.3**	<u>65.5</u>
HRNet	128×96	66.1	86.2	77.4	64.6	64.5
HRNet+DARK	128×96	69.6	87.1	76.2	67.0	72.9
HRNet(cp)+DARK	128×96	72.6	90.2	77.6	70.1	74.5
HRNet(cp)+ EHPE (Taylor)	128×96	**73.7**	**91.5**	**80.2**	**72.1**	**76.8**

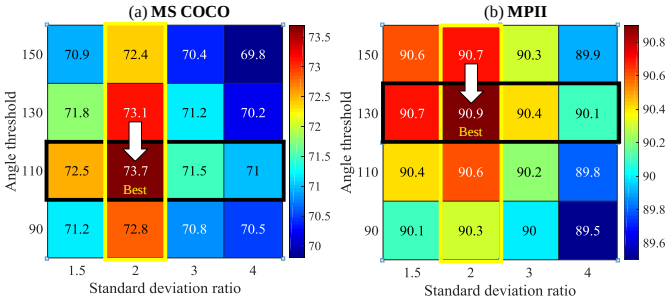

Fig. 7. Effect of standard deviation ratio σ_1/σ_2 and limb angle threshold T on model performance. (a) MS COCO database. (b) MPII database.

4)Standard deviation ratio and angle threshold

To solve the problem of model parameter selection, a series of investigations were conducted on the ratio of standard deviation of model performance and the limb angle threshold T on MS COCO and MPII, respectively. The former is the ratio of standard deviation of the x- and y-directions in the image coordinate system when the multi-Gaussian coordinate generation scheme was executed. The threshold value of the limb angle describes the degree of joint flexion. We used the method of control variables, setting the standard deviation ratio at [1.5, 2, 3, 4] while letting the limb angle threshold fluctuate between 90° and 150°. During this period, the other hyperparameters were kept unchanged. As shown in Fig. 7(a), when the standard deviation ratio equals 2 and the limb threshold is 110°, the model shows the optimal performance on the MSCOCO dataset. Meanwhile, in the MPII dataset indicated in Fig. 7(b), the model has the highest accuracy when the standard deviation ratio and extremity angle threshold are 2 (same as in MSCOCO) and 130°, respectively.

Second, the color changes of the two heatmaps reveals that the variation of the standard deviation ratio of multi-Gaussian coordinates has a great influence on the model effect of keypoint encoding. The standard deviation affects the distribution of the encoded multivariate Gaussian distribution. The standard deviation of the limb direction and that perpendicular to the limb direction should present a reasonable ratio range. On the one hand, if the ratio is too large, the model is likely to produce arbitrary judgment when neighboring interference occurs; on the other hand, when the ratio is too small, the network cannot adequately learn the structure of the human body. However, when the standard deviation ratio is fixed, the fluctuation of the edge angle threshold within a reasonable range has minimal effect on the performance of the model. The proposed coding scheme will achieve the maximum benefit when the threshold is between 110° and 130°.

5)Visualization of keypoint detections

We selected a human-dominated image that is challenging to recognize, and used a model equipped with optimal parameters to perform a positioning test of human keypoints. The visualized result is shown in Fig. 8. The human body is occluded by objects (such as in Figs. 8(d), 8(e), 8(i)), adjacent keypoint interference (Figs. 8(a), 8(f)), background confusions (such as Figs. 8(c), 8(f)), and other issues, and our positioning shows amazing robustness. Meanwhile, the above test experiments also concretely show that our model have the ability to solve the common interference problems in HPE tasks that cannot be ignored. Since the proposed EHPE model is not a lightweight architecture, it does not have much advantage at the time-consuming aspect while comparing with the state-of-the-art methods. For instance, the parsing takes 34ms for 9 people while the OpenPose [42] takes 0.58 ms. In future, we will introduce the knowledge distillation and pruning technologies to reduce inference time.

E. Expansion Experiments on the Infrared Images

In this section, an extended experiment based on the infrared images we captured is presented, as shown in Fig. 9. We gathered more than 20 volunteers and collected infrared images of 1,500 deputy classroom behavior, including individual and multi-person images. Infrared images have low resolution and low contrast with the absence of color and texture information compared with visible light-sufficient images. At the same time, their visual effects are blurred, and the grayscale distribution has a wireless relationship with the target reflection characteristics. However, infrared images are widely used in night vision, public security and other fields. The attitude estimation technology in infrared scenes is also in urgent need of development. Figures 9(a), 9(c), 9(d) and 9(e) are the original infrared images captured. The corresponding images show the test results estimated by our trained model. Experimental results show the proposed EHPE method can even work well in the infrared scenarios.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3197364

Fig. 8. Visualization results of human pose estimation in the light sufficient environment on MS COCO dataset.

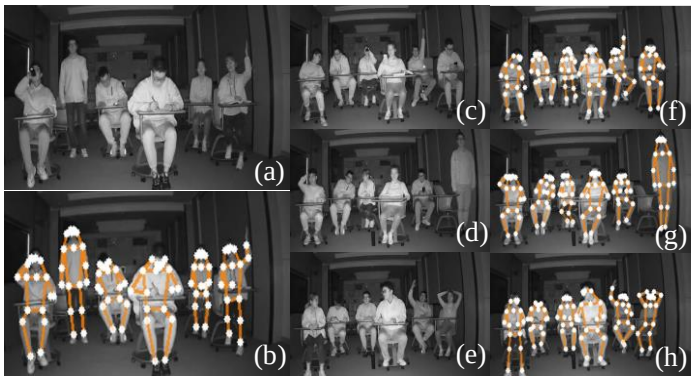

Fig. 9. HPE experimental results on the infrared images. (a),(c),(d),(e) Original HPE,(b),(f),(g),(h) Output image

V. .CONCLUSION

In this study, we propose an efficient human pose estimation model (EHPE) with skeleton cues-based Gaussian coordinate encoding. We consider extracting the relationship between adjacent joint points and describe them by the anisotropic Gaussian coordinate encoding. To the best of our knowledge, this study is the first to introduce the skeleton direction cues to heatmap distribution of HPE task. Then, the robust skeleton direction cues-aware architecture, which can learn the probability distribution we efficiently constructed and can explore keypoint information complementarity, is proposed. Furthermore, a multi-loss function is proposed to constrain the output to prevent overfitting. The *KL* divergence and Euclidean distance are selected to measure the predication label and ground truth one. We test EHPE on two HPE datasets. Experimental results demonstrate that EHPE can address the problems of ambiguity and occlusion in HPE and obtains a state-of-the-art performance compared with that of the existing methods. Furthermore, the success of EHPE demonstrates the importance of the skeleton direction cues in the HPE task, which is ignored by the previous researches.

REFERENCES

[1] A. Kamel, B. Sheng, P. Li, J. Kim, and D. D. Feng, "Hybrid Refinement-Correction Heatmaps for Human Pose Estimation," *IEEE Transactions on Multimedia*, vol. 23, pp. 1330 - 1342, 2021.

[2] M. Li, Z. Zhou, and X. Liu, "Multi-Person Pose Estimation Using Bounding Box Constraint and LSTM," *IEEE Transactions on Multimedia*, vol. 21, pp. 2653-2663, 2019.

[3] G. Ning, Z. Zhang, and Z. He, "Knowledge-Guided Deep Fractal Neural Networks for Human Pose Estimation," *IEEE Transactions on Multimedia*, vol. 20, pp. 1246-1259, 2018.

[4] Y.-L. Li, L. Xu, X. Liu, X. Huang, Y. Xu, S. Wang, *et al.*, "PaStaNet: Toward Human Activity Knowledge Engine," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 382-391.

[5] P. Baronti, M. Girolami, F. Mavilia, F. Palumbo, and G. Luisetto, "On the Analysis of Human Posture for Detecting Social Interactions with Wearable Devices," in *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*, 2020, pp. 1-6.

[6] Z. Luo, Z. Wang, Y. Huang, L. Wang, T. Tan, and E. Zhou, "Rethinking the Heatmap Regression for Bottom-up Human Pose Estimation," presented at the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR2021), 2021.

[7] H. Liu, H. Nie, Z. Zhang, and Y.-F. Li, "Anisotropic angle distribution learning for head pose estimation and attention understanding in human-computer interaction," *Neurocomputing*, vol. 433, pp. 310-322, 2021.

[8] W. Peng, X. Hong, H. Chen, and G. Zhao, "Learning Graph Convolutional Network for Skeleton-Based Human Action Recognition by Neural Searching," in *AAAI*, 2020, pp. 2669-2676.

[9] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," presented at *the Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference*, New Orleans, Louisiana, USA, 2018.

[10] W. Chen, H. Xu, C. Zhu, X. Liu, Y. Lu, C. Zheng, *et al.*, "Gaze Estimation via the Joint Modeling of Multiple Cues," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1390 - 1402, 2022.

[11] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, *et al.*, "Towards accurate multi-person pose estimation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4903-4911.

[12] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," in *European Conference on Computer Vision*, 2016, pp. 34-50.

[13] M. Ding and G. Fan, "Articulated and Generalized Gaussian Kernel Correlation for Human Pose Estimation," *IEEE Transactions on Image Processing*, vol. 25, pp. 776-789, 2016.

[14] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7103-7112.

[15] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2019, pp. 5693-5703.

[16] Z. Shao, Y. Li, and H. Zhang, "Learning Representations From Skeletal Self-Similarities for Cross-View Action Recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, pp. 160-174, 2021.

[17] J. Zhang, Z. Chen, and D. Tao, "Towards High Performance Human Keypoint Detection," International Journal of Computer Vision, vol. 129, no. 9, pp. 2639-2662, 2021.
[18] G. Borghi, M. Fabbri, R. Vezzani, S. Calderara, and R. Cucchiara, "Face-from-Depth for Head Pose Estimation on Depth Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 596-609, 2020.
[19] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1653-1660.
[20] J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," presented at the Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, Canada, 2014.
[21] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1831-1840.
[22] L. Ke, M.-C. Chang, H. Qi, and S. Lyu, "Multi-scale structure-aware network for human pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 713-728.
[23] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," in *Advances in neural information processing systems*, 2014, pp. 1799-1807.
[24] H. Fang, S. Xie, Y. Tai, and C. Lu, "RMPE: Regional Multi-person Pose Estimation," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2353-2362.
[25] J. Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, and C. Lu, "Crowdpose: Efficient crowded scenes pose estimation and a new benchmark," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10863-10872.
[26] X. Nie, J. Feng, Y. Zuo, and S. Yan, "Human pose estimation with parsing induced learner," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2100-2108.
[27] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 466-481.
[28] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European conference on computer vision*, 2016, pp. 483-499.
[29] H. Liu, T. Liu, Z. Zhang, A. K. Sangaiah, B. Yang, and Y. Li, "ARHPE: Asymmetric Relation-Aware Representation Learning for Head Pose Estimation in Industrial Human-Computer Interaction," *IEEE Transactions on Industrial Informatics*, vol. 18, pp. 7107 - 7117, 2022.
[30] B. Artacho and A. Savakis, "UniPose: Unified Human Pose Estimation in Single Images and Videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7035-7044.
[31] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, "Human pose estimation with iterative error feedback," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4733-4742.
[32] T. Liu, J. Wang, B. Yang, and X. Wang, "NGDNet: Nonuniform Gaussian-label distribution learning for infrared head pose estimation and on-task behavior understanding in the classroom," *Neurocomputing*, vol. 436, pp. 210-220, 2021.
[33] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2016, pp. 4724-4732.
[34] F. Zhang, X. Zhu, H. Dai, M. Ye, and C. Zhu, "Distribution-aware coordinate representation for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7093-7102.
[35] J. Huang, Z. Zhu, F. Guo, and G. Huang, "The Devil is in the Details: Delving into Unbiased Data Processing for Human Pose Estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5700-5709.
[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *presented at the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 2015*.
[37] O. Tuzel, F. Porikli, and P. Meer, "Region Covariance: A Fast Descriptor for Detection and Classification," in *European Conference on Computer Vision*, 2006.
[38] P. Li, J. Xie, Q. Wang, and W. Zuo, "Is second-order information helpful for large-scale visual recognition?," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2070-2078.
[39] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, *et al.*, "Microsoft COCO: Common Objects in Context," Cham, 2014, pp. 740-755.
[40] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D Human Pose Estimation: New Benchmark and State of the Art Analysis," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3686-3693.
[41] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961-2969.
[42] Z. Cao, G. Hidalgo, T. Simon *et al.*, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172 - 186, 2021.
[43] S. Huang, M. Gong, and D. Tao, "A Coarse-Fine Network for Keypoint Localization," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3047-3056.
[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[45] H. Liu, S. Fang, Z. Zhang, D. Li, K. Lin, and J. Wang, "MFDNet: Collaborative Poses Perception and Matrix Fisher Distribution for Head Pose Estimation," *IEEE Transactions on Multimedia*, vol. 24, pp. 2449 - 2460, 2022.

Hai LIU (Senior Member, IEEE) received the M.S. degree in applied mathematics from Huazhong University of Science and Technology (HUST), Wuhan, China, in 2010, and the Ph. D. degree in pattern recognition and artificial intelligence from the same university, in 2014.

Since June 2017, he has been an Assistant Professor with the National Engineering Research Center for E-Learning, Central China Normal University, Wuhan. From 2017 to 2019, he was a postdoctoral fellow in the Department of Mechanical Engineering, City University of Hong Kong, Kowloon, Hong Kong, where he was hosted by the Professor You-Fu Li. From 2020 to 2022, he was selected as "China- European Commission Talent Programme" under National Natural Science Foundation of China (NSFC). He was a senior researcher with UCL Interaction Centre, University College London, London, United Kingdom, where he was host by the Professor Sriram Subramanian. He has authored more than 70 peer reviewed articles in international journals from multiple domains. More than ten articles are selected as the ESI highly cited articles, and three paper was selected as the hot papers. His current research interests include deep learning, artificial intelligence, human pose estimation, gaze estimation, head pose estimation, educational technology and pattern recognition.

Dr. Liu has been frequently serving as a reviewer for more than six international journals including the *IEEE Transactions on Neural Networks and Learning Systems*, *IEEE Translations on Industrial Informatics*, *IEEE Translations on Cybernetics*, *IEEE Transactions on Knowledge and Data Engineering*. He is also a Communication Evaluation Expert for the NSFC from 2016 to present. He won the first prize of Science and Technology Progress Award by the Hubei Province of China in 2020.

Tingting Liu (Member IEEE) received the M.S. degree in Natural language processing from Huazhong University of Science and Technology, in 2014, Ph.D degree in education information technology from Central China Normal University (CCNU), Wuhan, China, in 2019.

She joined Hubei University, Wuhan, in 2020, and is currently an Assistant Professor with the School of Education. During Sep.2017-Sep. 2019, she was selected as a visiting scholar in School of Computer Science, Carnegie Mellon University, Pittsburgh, USA. From 2019 to 2020, she was a member of research staff with the Collaborative Innovation Centre for Information Technology and Balanced Development of K-12 Education, Faculty of Artificial Intelligence in Education, in CCNU, where she was hosted by the Professor Jixin Wang. Her current research interests include learning

behavior analysis, human pose estimation, label distribution learning and graph neural network.

Dr. Liu has been frequently serving as a Reviewer for several international journals including the *IEEE Transactions on Industrial Informatics*, *IEEE/ASME Transactions on Mechatronics*, *Neurocomputing*, and *International Journal of Human–Computer Interaction*. She is also a Communication Evaluation Expert for the National Natural Science Foundation of China from 2020.

Yu CHEN is currently working toward the M.S. degree in computer science and technology with the National Engineering Research Center for E-Learning, Central China Normal University, Wuhan, China. Her research interests include pattern recognition, human pose estimation, human-computer interaction and their applications in industry and big data analysis.

Zhaoli Zhang (Member, IEEE) received the M.S. degree in Computer Science from Central China Normal University, Wuhan, China, in 2004, and the Ph.D. degree in Computer Science from Huazhong University of Science and Technology in 2008.

He is currently a professor in the National Engineering Research Center for e-learning, Central China Normal University, Wuhan, China. His research interests include self-regulated learning, human-computer interaction, deep learning, image processing, knowledge services and software engineering. He is a member of IEEE and CCF (China Computer Federation).

Youfu Li (Fellow, IEEE) received the B.S. and M.S. degrees in electrical engineering from the Harbin Institute of Technology, Harbin, China, and the Ph.D. degree in robotics from the Department of Engineering Science, University of Oxford, Oxford, U.K., in 1993.

From 1993 to 1995, he was a Research Staff with the Department of Computer Science, University of Wales, Aberystwyth, U.K. He joined the City University of Hong Kong, Hong Kong, in 1995, and is currently a Professor with the Department of Mechanical Engineering. His current research interests include robot sensing, robot vision, 3D vision, and visual tracking.

Dr. Li has served as an Associate Editor for the IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING and is currently serving as an Associate Editor for the IEEE Robotics and Automation Magazine. He is an Editor of the IEEE Robotics and Automation Society's Conference Editorial Board and the IEEE Conference on Robotics and Automation.