A Characteristic Function-based Method for Bottom-up Human Pose Estimation

Haoxuan Qu[1], Yujun Cai[2], Lin Geng Foo[1], Ajay Kumar[3], Jun Liu [1,*]
[1]Singapore University of Technology and Design, Singapore
[2]Nanyang Technological University, Singapore
[3]The Hong Kong Polytechnic University, Hong Kong
haoxuan_qu@mymail.sutd.edu.sg, yujun001@e.ntu.edu.sg, lingeng_foo@mymail.sutd.edu.sg
ajay.kumar@polyu.edu.hk, jun_liu@sutd.edu.sg

Abstract

Most recent methods formulate the task of human pose estimation as a heatmap estimation problem, and use the overall L2 loss computed from the entire heatmap to optimize the heatmap prediction. In this paper, we show that in bottom-up human pose estimation where each heatmap often contains multiple body joints, using the overall L2 loss to optimize the heatmap prediction may not be the optimal choice. This is because, minimizing the overall L2 loss cannot always lead the model to locate all the body joints across different sub-regions of the heatmap more accurately. To cope with this problem, from a novel perspective, we propose a new bottom-up human pose estimation method that optimizes the heatmap prediction via minimizing the distance between two characteristic functions respectively constructed from the predicted heatmap and the groundtruth heatmap. Our analysis presented in this paper indicates that the distance between these two characteristic functions is essentially the upper bound of the L2 losses w.r.t. sub-regions of the predicted heatmap. Therefore, via minimizing the distance between the two characteristic functions, we can optimize the model to provide a more accurate localization result for the body joints in different sub-regions of the predicted heatmap. We show the effectiveness of our proposed method through extensive experiments on the COCO dataset and the CrowdPose dataset.

1. Introduction

Human pose estimation aims to locate the body joints of each person in a given RGB image. It is relevant to various applications, such as action recognition [7, 43], person Re-ID [28], and human object interaction [35]. For tackling human pose estimation, most of the recent methods fall into two major categories: *top-down* methods and *bottom-up* methods. *Top-down* methods [24,32,33,39,44] generally use a human detector to detect all the people in the image, and then perform single-person pose estimation for each detected subject separately. In contrast, *bottom-up* methods [5,6,16,17,22,23,25,26] usually locate the body joints of all people in the image at the same time. Hence, *bottom-up* methods, the main focus of this paper, are often a more efficient choice compared to *top-down* methods, especially when there are many people in the input image [5].

In existing works, it is common to regard human pose estimation as a heatmap prediction problem, since this can preserve the spatial structure of the input image throughout the encoding and decoding process [12]. During the general optimization process, the groundtruth (GT) heatmaps $\mathbf{H}_g$ are first constructed via putting 2D Gaussian blobs centered at the GT coordinates of the body joints. After that, these constructed GT heatmaps are used to supervise the predicted heatmaps $\mathbf{H}_p$ via the overall L2 loss $L_2^{overall}$ calculated (averaged) over the whole heatmap. Specifically, denoting the area of the heatmap as A, we have $L_2^{overall} = \frac{\|\mathbf{H}_p - \mathbf{H}_g\|_2^2}{A}$.

We argue that using the overall L2 loss to supervise the predicted heatmap may not be the optimal choice in bottom-up methods where each heatmap often contains multiple body joints from the multiple people in various sub-regions, as shown in Fig. 1(b). This is because, a smaller overall L2 loss calculated over the whole heatmap cannot always lead the model to locate all the body joints across different sub-regions in the heatmap more accurately. As illustrated in Fig. 1(a), the predicted heatmap #2 has a smaller overall L2 loss compared to the predicted heatmap #1. However, the predicted heatmap #2 **locates the body joint in the top-right sub-region wrongly**, whereas the predicted heatmap #1 **locates body joints in both the top-right and bottom-left sub-regions correctly**. This is because, while the decrease of the overall L2 loss can be achieved when the L2 loss w.r.t. each sub-region either decreases or remains the

*Corresponding Author

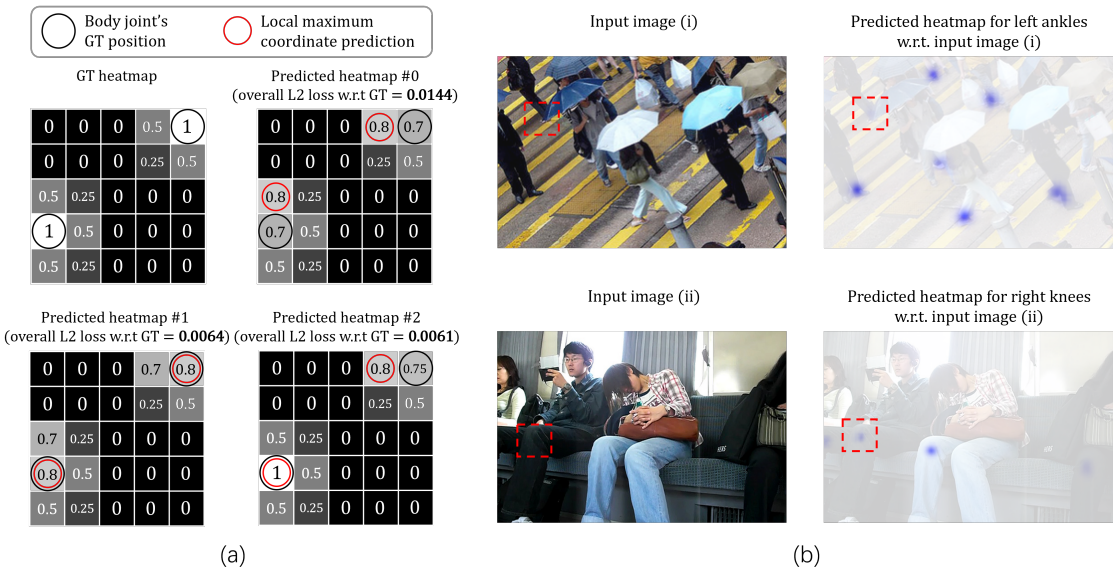

Figure 1. (a) Illustration of heatmaps. The predicted heatmap #2 with a smaller overall L2 loss locates the body joint in the top-right sub-region wrongly, while the predicted heatmap #1 with a larger overall L2 loss locates body joints in both the top-right and bottom-left sub-regions correctly. (b) Output of a commonly used bottom-up method, HrHRNet-W32 [6]. As shown, it misses left ankle in the dashed sub-region of image (i) completely, and misidentifies right knee in the dashed sub-region of image (ii). This indicates that accurately localizing the body joints of multiple people in a single heatmap is a challenging problem. (Best viewed in color.)

same (e.g., from predicted heatmap #0 to predicted heatmap #1), it can also be achieved when there is a decrease of L2 loss w.r.t. certain sub-regions and an increase of L2 loss for some other sub-regions (e.g., from predicted heatmap #1 to predicted heatmap #2). This indicates that, in bottom-up methods, the decrease of the overall L2 loss does not always lead to a more accurate localization result for the body joints in different sub-regions of the predicted heatmap at the same time. Besides, we also show some results of a commonly used bottom-up method, HrHRNet-W32 [6], in Fig. 1(b). As shown, it may even miss or misidentify certain body joints when there are a number of people in the input image. This indicates that it is quite difficult to accurately locate all body joints of all people in the predicted heatmap.

To tackle the above-mentioned problem in bottom-up methods, in this paper, rather than using the overall L2 loss to supervise the whole heatmap, we instead aim to *optimize the body joints over sub-regions of the predicted heatmap at the same time*. To this end, from a new perspective, we express the predicted and GT heatmaps as characteristic functions, and minimize the difference between these functions, allowing different sub-regions of the predicted heatmap to be optimized at the same time.

More specifically, we first construct two distributions respectively from the predicted heatmap and the GT heatmap. After that, we obtain two characteristic functions of these two distributions and optimize the heatmap prediction via minimizing the distance between these two characteristic functions. We analyze in Sec. 3.3 that the distance between the two characteristic functions is the upper bound of the L2 losses w.r.t sub-regions in the predicted heatmap. Therefore, via minimizing the distance between the two characteristic functions, our method can locate body joints in different sub-regions more accurately at the same time, and thus achieve superior performance.

The **contributions** of our work are summarized as follows. 1) From a new perspective, we supervise the predicted heatmap using the distance between the characteristic functions of the predicted and GT heatmaps. 2) We analyze (in Sec. 3.3) that the L2 losses w.r.t. sub-regions of the predicted heatmap are upper-bounded by the distance between the characteristic functions. 3) Our proposed method achieves state-of-the-art performance on the evaluation benchmarks [19, 21].

2. Related Work

Human Pose Estimation. Due to the wide range of applications, human pose estimation has received lots of attention [5, 6, 16, 17, 22–26, 29, 32, 33, 39, 44], and most of the recent methods fall into two categories: *top-down* methods and *bottom-up* methods. In *top-down* methods, a human detector is generally used to detect all the people in the image first, and then single-person pose estimation is conducted for each detected subject separately. The single-person pose estimation methods that are commonly used in *top-down* methods include Hourglass [24], Simple Baseline [39], HRNet [32], and HRFormer [44], etc. Besides *top-down* methods, *bottom-up* methods [5, 6, 16, 17, 22, 23, 25, 26] have also attracted a lot of attention recently due to its efficiency [5].

In *bottom-up* methods, most methods first detect all

identity-free body joints over the whole input image, and then group them into different people. Among these methods, DeepCut and Person-Lab [14,26,27] incorporate offset fields into their methods, while Openpose and PifPaf [5,17] make use of part affinity fields in their methods. From another perspective, associate embedding [23] teaches the model to output the group assignments and the localization results of the body joints at the same time, and HGG [16] further combines graph neural networks on top of the associate embedding. Besides the above methods, there also exist some *bottom-up* methods [11,25,45] that directly regress the coordinates of body joints belonging to the same person.

Existing heatmap-based bottom-up methods often use an overall L2 loss calculated over the whole heatmap to optimize heatmap prediction. Differently, in this paper, we propose a new bottom-up method that optimizes the heatmap prediction via minimizing the difference between the characteristic functions of the predicted and GT heatmaps.

Characteristic Function. The characteristic function, a concept originally proposed in probability theory and statistics, has been studied in various areas [2,8,9,13,20,31,41] over the years, such as two-sample testing [8,9,13], generative adversarial nets [2,20], and few-shot classification [41]. Inspired by these works, in this paper, from a novel perspective, we propose to optimize the heatmap prediction for bottom-up human pose estimation via minimizing the distance between two characteristic functions. We theoretically analyze that the distance between the two characteristic functions respectively constructed from the predicted heatmap and the GT heatmap is the upper bound of the L2 losses w.r.t. sub-regions of the predicted heatmap.

3. Method

In bottom-up human pose estimation, as shown in Fig. 1(a), minimizing the overall L2 loss between the predicted heatmap and the GT heatmap cannot always lead the model to locate all the body joints across different sub-regions of the heatmap more accurately. In this work, we aim to optimize the body joints over sub-regions of the predicted heatmap at the same time. To achieve this, we propose a new bottom-up method that optimizes the heatmap prediction via minimizing the distance between two characteristic functions constructed from the predicted and GT heatmaps.

Below, we first briefly introduce the characteristic function, and then discuss how we formulate the heatmap optimization process. After that, we show the theoretical analysis of our proposed method.

3.1. Revisiting Characteristic Function

The characteristic function is generally used in probability theory and statistics. Given an N-dimensional distribution D, its corresponding characteristic function φ_D can be written as:

$$\varphi_D(\mathbf{t}) = E_{\mathbf{x}\sim D}[e^{i\langle \mathbf{t},\mathbf{x}\rangle}] = \int_{\mathbb{R}^N} e^{i\langle \mathbf{t},\mathbf{x}\rangle}\, dD \tag{1}$$

where E represents expectation, $i^2 = -1$, $\langle\cdot,\cdot\rangle$ represents dot product, $\mathbf{t}$ is a random N-dimensional vector, and $\mathbf{x}$ is an N-dimensional vector sampled from D. Note that the characteristic function always exists and has a one-to-one correspondence with the distribution. Besides, the characteristic function is the Fourier transform of the probability density function if the latter exists as well. Moreover, the characteristic function is always finite and bounded ($|\varphi_D(\mathbf{t})| \leq 1$). This makes calculation of the distance between two characteristic functions always meaningful.

3.2. Proposed Heatmap Optimization Process

Below, we discuss how we formulate the heatmap optimization process for bottom-up human pose estimation via (1) constructing two distributions from the predicted heatmap and the GT heatmap respectively; (2) calculating characteristic functions from these two distributions; and (3) formulating the loss function as the distance between the two characteristic functions.

Distribution Construction. Given an input image, for each type of body joints, we denote the corresponding predicted heatmap as H_p and the corresponding GT heatmap as H_g. We propose to formulate the two distributions $D(H_p)$ and $D(H_g)$ from the two heatmaps H_p and H_g with the following two steps. (1) As distributions cannot hold negative probabilities, we first pass H_p through a $relu$ activation function to make it non-negative. Note that H_g is already non-negative. (2) After that, as the sum of probabilities of each constructed distribution needs to be 1, we further normalize both the output of step (1) and H_g. Hence, with the above two steps, we formulate $D(H_p)$ and $D(H_g)$ as:

$$D(H_p) = \frac{relu(H_p)}{\|relu(H_p)\|_1}, \quad D(H_g) = \frac{H_g}{\|H_g\|_1} \tag{2}$$

Characteristic Function Calculation. For every type of body joints, after formulating the two distributions $D(H_p)$ and $D(H_g)$, we follow Eq. 1 to calculate the two characteristic functions $\varphi_{D(H_p)}(\mathbf{t})$ and $\varphi_{D(H_g)}(\mathbf{t})$ as:

$$\begin{aligned} \varphi_{D(H_p)}(\mathbf{t}) &= E_{\mathbf{x}\sim D(H_p)}[e^{i\langle \mathbf{t},\mathbf{x}\rangle}], \\ \varphi_{D(H_g)}(\mathbf{t}) &= E_{\mathbf{x}\sim D(H_g)}[e^{i\langle \mathbf{t},\mathbf{x}\rangle}] \end{aligned} \tag{3}$$

Loss Function Formulation. Above we discuss how we obtain the two characteristic functions w.r.t. the predicted heatmap and the GT heatmap for a single type of body joints. Note that in bottom-up human pose estimation, multiple types of body joints are required to be located at the same time. Here, we first discuss how we formulate the loss function for a single type of body joints, and then introduce the overall loss function for all types of body joints.

To formulate the loss function for the k-th type of body

joints, given the two characteristic functions $\varphi^k_{D(H_p)}(\mathbf{t})$ and $\varphi^k_{D(H_g)}(\mathbf{t})$, we first write the loss function L_k as the distance between these two characteristic functions [2]:

$$L_k = \int_{\mathbb{R}^2} \|\varphi^k_{D(H_p)}(\mathbf{t}) - \varphi^k_{D(H_g)}(\mathbf{t})\|_2^2 \omega(\mathbf{t},\eta)\, d\mathbf{t} \quad (4)$$

where $\omega(\mathbf{t},\eta)$ is a weighting function. Here we set $\omega(\mathbf{t},\eta)$ to be the probability density function of a uniform distribution in B_U, where $B_U = [-U,U]\times[-U,U]$ is a finite predefined range and U is a hyperparameter. This means that, $\omega(\mathbf{t},\eta) = \frac{1}{4U^2}$ when $\mathbf{t} \in B_U$ and $\omega(\mathbf{t},\eta) = 0$ otherwise. We thus further rewrite Eq. 4 as:

$$L_k = \int_{B_U} \|\frac{1}{2U}\big(\varphi^k_{D(H_p)}(\mathbf{t}) - \varphi^k_{D(H_g)}(\mathbf{t})\big)\|_2^2\, d\mathbf{t} \quad (5)$$

Finally, from Eq. 5, we formulate the loss function L_k as:

$$L_k = \int_{B_U} \|\frac{\gamma}{2U}\big(\varphi^k_{D(H_p)}(\mathbf{t}) - \varphi^k_{D(H_g)}(\mathbf{t})\big)\|_2^2\, d\mathbf{t} \quad (6)$$

where $\gamma = \frac{U^2\sqrt{A}}{\pi^2}$ is a constant coefficient and A is the area of the heatmap. Note that Eq. 6 is equivalent to Eq. 5 during the optimization process, as the efficacy of the added constant γ can be achieved by adjusting the learning rate.

After getting the loss function for each type of body joints, we formulate the total loss for all types of joints as:

$$L_{total} = \sum_{k=1}^{K} L_k \quad (7)$$

where K denotes the total number of body joint types.

3.3. Theoretical Analysis

Below, we perform theoretical analysis to show the effectiveness of our method for bottom-up human pose estimation. Before going into the theorem, we first introduce a lemma that can facilitate the proof of the theorem.

Lemma 1. *Let φ_D be the characteristic function of a 2-dimensional distribution D. Let $R^r = [x_1^{lower}, x_1^{upper}] \times [x_2^{lower}, x_2^{upper}]$ a rectangular region, $R^e = \{x_1^{lower}, x_1^{upper}\} \times [x_2^{lower}, x_2^{upper}] \cup [x_1^{lower}, x_1^{upper}] \times \{x_2^{lower}, x_2^{upper}\}$ the edges of this region, and $R^v = \{x_1^{lower}, x_1^{upper}\} \times \{x_2^{lower}, x_2^{upper}\}$ the vertices of this region. Let $B_T = [-T,T]\times[-T,T]$. Denote $[D]_R$ the portion of the distribution D in R. $[D]_{R^r}$ can then be written as:*

$$[D]_{R^r} = \Bigg(\lim_{T\to\infty} \frac{1}{(2\pi)^2} \int_{B_T} \Big(\prod_{n=1}^{2} \big(\frac{e^{-it_n x_n^{lower}} - e^{-it_n x_n^{upper}}}{it_n}\big)\Big) \varphi_D(\mathbf{t})\Big)\, dt_1 dt_2 \Bigg) + \epsilon([D]_{R^r}) \quad (8)$$

where $\epsilon([D]_{R^r}) = \frac{[D]_{R^e}}{2} + \frac{[D]_{R^v}}{4}$ and $dt_1 dt_2$ are calculated based on the Lebesgue measure.

The proof of Lemma 1 is provided in the supplementary. After introducing this lemma, we analyze our proposed method below.

Theorem 1. *Let R^r_{sub} be a random rectangular sub-region in the heatmap of the k-th type of body joints where $\|[D(H_p)]_{R^e_{sub}} - [D(H_g)]_{R^e_{sub}}\|_2^2$ is relatively small compared to $\|[D(H_p)]_{R^r_{sub}} - [D(H_g)]_{R^r_{sub}}\|_2^2$. The relation between the L2 loss w.r.t. this sub-region and L_k can be written as:*

$$\frac{\|[D(H_p)]_{R^r_{sub}} - [D(H_g)]_{R^r_{sub}}\|_2^2}{\lambda(R^r_{sub})} \le L_k \quad (9)$$

Note that $\lambda(R^r_{sub})$ as the Lebesgue measure represents the area of R^r_{sub}.

Proof. To prove Theorem 1, we first reformulate Lemma 1 as:

$$\begin{aligned}[D]_{R^r} &= \Bigg(\lim_{T\to\infty} \frac{1}{(2\pi)^2} \int_{B_T} \Big(\prod_{n=1}^{2} \big(\frac{e^{-it_n x_n^{lower}} - e^{-it_n x_n^{upper}}}{it_n}\big)\Big) \varphi_D(\mathbf{t})\Big)\, dt_1 dt_2 \Bigg) + \epsilon([D]_{R^r}) \\ &= \lim_{T\to\infty} \frac{1}{(2\pi)^2} \int_{B_T} \varphi_D(\mathbf{t}) \int_{R^r} e^{-i\langle \mathbf{t},\mathbf{x}\rangle}\, d\mathbf{x}\ d\mathbf{t}\ + \epsilon([D]_{R^r}) \end{aligned} \quad (10)$$

where $d\mathbf{t} = dt_1 dt_2$, and both $d\mathbf{x}$ and $d\mathbf{t}$ are calculated based on the Lebesgue measure.
After that, we rewrite $\|[D(H_p)]_{R^r_{sub}} - [D(H_g)]_{R^r_{sub}}\|_2^2$ as:

$$\begin{aligned} &\|[D(H_p)]_{R^r_{sub}} - [D(H_g)]_{R^r_{sub}}\|_2^2 && (11)\\ &\approx \|[D(H_p)]_{R^r_{sub}} - [D(H_g)]_{R^r_{sub}} && (12)\\ &\quad - \big(\epsilon([D(H_p)]_{R^r_{sub}}) - \epsilon([D(H_g)]_{R^r_{sub}})\big)\|_2^2 \\ &= \| \lim_{T\to\infty} \frac{1}{(2\pi)^2} \int_{B_T} \varphi^k_{D(H_p)}(\mathbf{t}) \int_{R^r_{sub}} e^{-i\langle \mathbf{t},\mathbf{x}\rangle}\, d\mathbf{x}\ d\mathbf{t} && (13)\\ &\quad - \lim_{T\to\infty} \frac{1}{(2\pi)^2} \int_{B_T} \varphi^k_{D(H_g)}(\mathbf{t}) \int_{R^r_{sub}} e^{-i\langle \mathbf{t},\mathbf{x}\rangle}\, d\mathbf{x}\ d\mathbf{t}\ \|_2^2 \\ &= \| \lim_{T\to\infty} \int_{B_T} \int_{R^r_{sub}} \frac{\varphi^k_{D(H_p)}(\mathbf{t}) - \varphi^k_{D(H_g)}(\mathbf{t})}{(2\pi)^2} e^{-i\langle \mathbf{t},\mathbf{x}\rangle}\, d\mathbf{x}\ d\mathbf{t}\ \|_2^2 && (14)\\ &\approx \| \int_{B_U} \int_{R^r_{sub}} \frac{\varphi^k_{D(H_p)}(\mathbf{t}) - \varphi^k_{D(H_g)}(\mathbf{t})}{(2\pi)^2} e^{-i\langle \mathbf{t},\mathbf{x}\rangle}\, d\mathbf{x}\ d\mathbf{t}\ \|_2^2 && (15)\\ &\le 4U^2 A \int_{B_U} \int_{R^r_{sub}} \| \frac{\varphi^k_{D(H_p)}(\mathbf{t}) - \varphi^k_{D(H_g)}(\mathbf{t})}{(2\pi)^2} e^{-i\langle \mathbf{t},\mathbf{x}\rangle} \|_2^2\, d\mathbf{x}\ d\mathbf{t} && (16) \end{aligned}$$

Table 1. Comparisons with bottom-up methods on the **COCO val2017** set (single-scale testing).

Method	Venue	Backbone	Input size	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
OpenPose [5]	CVPR 2017	VGG-19	-	61.0	84.9	67.5	56.3	69.3
HGG [16]	ECCV 2020	Hourglass	512	60.4	83.0	66.2	-	-
PersonLab [26]	ECCV 2018	ResNet-152	1401	66.5	86.2	71.9	62.3	73.2
PifPaf [17]	CVPR 2019	ResNet-152	-	67.4	-	-	-	-
PETR [30]	CVPR 2022	-	1333	67.4	87.0	74.9	61.7	75.9
DEKR [11]	CVPR 2021	HRNet-W48	640	71.0	88.3	77.4	66.7	78.5
PINet [37]	NIPS 2021	HRNet-W32	512	67.4	-	-	-	-
CIR&QEM [40]	AAAI 2022	HRNet-W48	640	72.4	89.1	-	67.3	80.4
CID [36]	CVPR 2022	HRNet-W32	512	66.0	86.7	72.3	59.8	76.0
LOGP-CAP [42]	CVPR 2022	HRNet-W48	640	72.2	88.9	78.9	68.1	78.9
SWAHR [22]	CVPR 2021	HrHRNet-W32	512	68.9	87.8	74.9	63.0	77.4
SWAHR [22]	CVPR 2021	HrHRNet-W48	640	70.8	88.5	76.8	66.3	77.4
CenterAttention [4]	ICCV 2021	HrHRNet-W32	512	68.6	87.6	74.1	62.0	78.0
PoseTrans [15]	ECCV 2022	HrHRNet-W32	512	68.4	87.1	74.8	62.7	77.1
HrHRNet [6]	CVPR 2020	HrHRNet-W32	512	67.1	86.2	73.0	61.5	76.1
+ Ours		HrHRNet-W32	512	**69.9(↑2.8)**	88.1	76.0	64.2	78.1
HrHRNet [6]	CVPR 2020	HrHRNet-W48	640	69.9	87.2	76.1	65.4	76.4
+ Ours		HrHRNet-W48	640	**72.5(↑2.6)**	89.3	79.1	68.3	79.0

Table 2. Comparisons with bottom-up methods on the **COCO val2017** set (multi-scale testing).

Method	Venue	Backbone	Input size	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
HGG [16]	ECCV 2020	Hourglass	512	68.3	86.7	75.8	-	-
Point-Set Anchors [38]	ECCV 2020	HRNet-W48	640	69.8	88.8	76.3	-	-
DEKR [11]	CVPR 2021	HRNet-W48	640	72.3	88.3	78.6	68.6	78.6
SWAHR [22]	CVPR 2021	HrHRNet-W32	512	71.4	88.9	77.8	66.3	78.9
SWAHR [22]	CVPR 2021	HrHRNet-W48	640	73.2	89.8	79.1	69.1	79.3
PoseTrans [15]	ECCV 2022	HrHRNet-W32	512	71.2	88.2	77.2	66.5	78.0
HrHRNet [6]	CVPR 2020	HrHRNet-W32	512	69.9	87.1	76.0	65.3	77.0
+ Ours		HrHRNet-W32	512	**71.8(↑1.9)**	88.9	78.1	67.3	78.4
HrHRNet [6]	CVPR 2020	HrHRNet-W48	640	72.1	88.4	78.2	67.8	78.3
+ Ours		HrHRNet-W48	640	**73.7(↑1.6)**	89.9	79.6	69.6	79.5

$$\leq 4U^2A \int_{B_U} \int_{R^r_{sub}} \| \frac{\varphi^k_{D(H_p)}(\mathbf{t}) - \varphi^k_{D(H_g)}(\mathbf{t})}{(2\pi)^2} \|_2^2 \, d\mathbf{x} \; d\mathbf{t} \tag{17}$$

$$= 4U^2A \int_{R^r_{sub}} \int_{B_U} \| \frac{\varphi^k_{D(H_p)}(\mathbf{t}) - \varphi^k_{D(H_g)}(\mathbf{t})}{(2\pi)^2} \|_2^2 \, d\mathbf{t} \; d\mathbf{x} \tag{18}$$

$$= \int_{R^r_{sub}} \int_{B_U} \| \frac{\gamma}{2U} (\varphi^k_{D(H_p)}(\mathbf{t}) - \varphi^k_{D(H_g)}(\mathbf{t})) \|_2^2 \, d\mathbf{t} \; d\mathbf{x} \tag{19}$$

$$= L_k \lambda(R^r_{sub}) \tag{20}$$

where Eq. 12 holds since $\|[D(H_p)]_{R^e_{sub}} - [D(H_g)]_{R^e_{sub}}\|_2^2$ is relatively small compared to $\|[D(H_p)]_{R^r_{sub}} - [D(H_g)]_{R^r_{sub}}\|_2^2$, Eq. 13 holds because of Eq. 10, Eq. 15 holds based on the analysis in the supplementary, Eq. 16 holds due to the continuity of L2 distance and the Cauchy–Schwarz inequality, Eq. 17 holds due to the fact that $\|e^{-i\langle \mathbf{t}, \mathbf{x} \rangle}\|_2^2 = 1$ and the Cauchy–Schwarz inequality, Eq. 18 holds due to Fubini's theorem.

We can then move $\lambda(R^r_{sub})$ on the right hand side of Eq. 20 to the left hand side to get Theorem 1. □

As shown in Theorem 1, for the sub-region R^r_{sub}, when the sum of the pixelwise L2 distances between the predicted and GT heatmaps over this entire sub-region is relatively large compared to only over its edges, L_k will be the upper bound of the L2 loss w.r.t. this sub-region. Because of this, via minimizing L_k, we can enable the L2 losses w.r.t. all such sub-regions to be smaller. Note that such sub-regions can be easily found, since the edge of a sub-region typically contains many less pixels compared to the entire sub-region in the first place. Furthermore, for sub-regions containing missed or inaccurate body joints in its center, which are precisely the erroneous predictions that need to be corrected, the sum of the pixelwise L2 distances over the entire sub-region will then be much larger compared to only over its edge. Therefore, our method can optimize the model to provide a more accurate localization result for the body joints in different sub-regions of the predicted heatmap at the same time, whereas the existing bottom-up methods, usually relying on the overall L2 loss, do not hold this property. Thus, our method can achieve superior performance for bottom-up human pose estimation.

Note that during implementation, since L_k itself as an integral is not tractable, inspired by [2], we define $\hat{L}_k$ as a

Table 3. Comparisons with bottom-up methods on the **COCO test-dev2017** set (single-scale testing).

Method	Venue	Backbone	Input size	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
OpenPose [5]	CVPR 2017	VGG-19	-	61.8	84.9	67.5	57.1	68.2
Hourglass [24]	ECCV 2016	Hourglass	512	56.6	81.8	61.8	49.8	67.0
Associative Embedding [23]	NIPS 2017	Hourglass	512	56.6	81.8	61.8	49.8	67.0
SPM [25]	ICCV 2019	Hourglass	-	66.9	88.5	72.9	62.6	73.1
MDN [34]	CVPR 2020	Hourglass	-	62.9	85.1	69.4	58.8	71.4
PersonLab [26]	ECCV 2018	ResNet-152	1401	66.5	88.0	72.6	62.4	72.3
PifPaf [17]	CVPR 2019	ResNet-152	-	66.7	-	-	62.4	72.9
PETR [30]	CVPR 2022	SWin-L	1333	70.5	91.5	78.7	65.2	78.0
DEKR [11]	CVPR 2021	HRNet-W48	640	70.0	89.4	77.3	65.7	76.9
PINet [37]	NIPS 2021	HRNet-W32	512	66.7	-	-	-	-
CIR&QEM [40]	AAAI 2022	HRNet-W48	640	71.0	90.2	78.2	66.2	77.8
CID [36]	CVPR 2022	HRNet-W48	640	70.7	90.3	77.9	66.3	77.8
LOGP-CAP [42]	CVPR 2022	HRNet-W48	640	70.8	89.7	77.8	66.7	77.0
SWAHR [22]	CVPR 2021	HrHRNet-W48	640	70.2	89.9	76.9	65.2	77.0
CenterAttention [4]	ICCV 2021	HrHRNet-W48	640	69.6	89.7	76.0	64.9	76.3
PoseTrans [15]	ECCV 2022	HrHRNet-W32	512	67.4	88.3	73.9	62.1	75.1
HrHRNet [6]	CVPR 2020	HrHRNet-W32	512	66.4	87.5	72.8	61.2	74.2
+ Ours		HrHRNet-W32	512	**68.9(↑2.5)**	89.2	75.7	63.7	76.1
HrHRNet [6]	CVPR 2020	HrHRNet-W48	640	68.4	88.2	75.1	64.4	74.2
+ Ours		HrHRNet-W48	640	**71.1(↑2.7)**	90.4	78.2	66.9	77.2

Table 4. Comparisons with bottom-up methods on the **COCO test-dev2017** set (multi-scale testing).

Method	Venue	Backbone	Input size	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
Hourglass [24]	ECCV 2016	Hourglass	512	63.0	85.7	68.9	58.0	70.4
Associative Embedding [23]	NIPS 2017	Hourglass	512	63.0	85.7	68.9	58.0	70.4
HGG [16]	ECCV 2020	Hourglass	512	67.6	85.1	73.7	62.7	74.6
SimplePose [18]	AAAI 2020	IMHN	512	68.1	-	-	66.8	70.5
PersonLab [26]	ECCV 2018	-	1401	68.7	89.0	75.4	64.1	75.5
PETR [30]	CVPR 2022	SWin-L	1333	71.2	91.4	79.6	66.9	78.0
Point-Set Anchors [38]	ECCV 2020	HRNet-W48	640	68.7	89.9	76.3	64.8	75.3
DEKR [11]	CVPR 2021	HRNet-W48	640	71.0	89.2	78.0	67.1	76.9
CIR&QEM [40]	AAAI 2022	HRNet-W48	640	71.7	90.4	78.7	67.3	78.5
SWAHR [22]	CVPR 2021	HrHRNet-W48	640	72.0	90.7	78.8	67.8	77.7
CenterAttention [4]	ICCV 2021	HrHRNet-W48	640	71.1	90.5	77.5	66.9	76.7
PoseTrans [15]	ECCV 2022	HrHRNet-W32	512	69.9	89.3	77.0	65.2	76.2
HrHRNet [6]	CVPR 2020	HrHRNet-W32	512	69.0	89.0	75.8	64.4	75.2
+ Ours		HrHRNet-W32	512	**70.8(↑1.8)**	90.1	77.8	66.0	77.3
HrHRNet [6]	CVPR 2020	HrHRNet-W48	640	70.5	89.3	77.2	66.6	75.8
+ Ours		HrHRNet-W48	640	**72.3(↑1.8)**	91.5	79.8	67.9	78.2

tractable alternative of L_k as:

$$\hat{L}_k = \sum_{m=1}^{M} \| \frac{\gamma}{2U} \left(\varphi^k_{D(H_p)}(\mathbf{t}_m) - \varphi^k_{D(H_g)}(\mathbf{t}_m) \right) \|_2^2 \quad (21)$$

where $\{\mathbf{t}_1, ..., \mathbf{t}_M\}$ denotes a set of M vectors randomly sampled from B_U.

The total loss $\hat{L}_{total}$ for all body joint types can then be written as:

$$\hat{L}_{total} = \sum_{k=1}^{K} \hat{L}_k \quad (22)$$

3.4. Overall Training and Testing

Here we discuss the overall training and testing scheme of our method. Specifically, during training, we supervise the predicted heatmaps via the total loss in Eq. 22 instead of using the commonly used overall L2 loss, and following [6, 22, 23], we conduct grouping via associate embedding. During testing, we follow the evaluation procedure of previous works [6, 22] that conduct bottom-up human pose estimation. Note that in experiments, it is easy to implement $\hat{L}_k$ in Eq. 21, and we provide more details on how we implement $\hat{L}_k$ in experiments in the supplementary.

4. Experiments

To evaluate the effectiveness of our method for bottom-up human pose estimation, we conduct experiments on the COCO dataset [21] and the CrowdPose dataset [19]. Besides, we also test the effectiveness of our method on top-down methods in the supplementary. We conduct our experiments on RTX 3090 GPUs.

4.1. COCO Keypoint Detection

Dataset & evaluation metric. The COCO dataset [21] contains over 200k images, and in this dataset, each person instance is annotated with 17 body joints. This dataset consists of three subsets including COCO training set (57k

Table 5. Comparisons with bottom-up methods on the **CrowdPose testing** set.

Method	Venue	Backbone	Input size	AP	AP^{50}	AP^{75}	AP^{E}	AP^{M}	AP^{H}
w/ single-scale testing									
OpenPose [5]	CVPR 2017	VGG-19	-	-	-	-	62.7	48.7	32.3
HrHRNet [6]	CVPR 2020	HrHRNet-W48	640	65.9	86.4	70.6	73.3	66.5	57.9
PETR [30]	CVPR 2022	-	-	72.0	90.9	78.8	78.0	72.5	65.4
DEKR [11]	CVPR 2021	HRNet-W48	640	67.3	86.4	72.2	74.6	68.1	58.7
PINet [37]	NIPS 2021	HRNet-W32	512	68.9	88.7	74.7	75.4	69.6	61.5
CID [36]	CVPR 2022	HRNet-W48	640	72.3	90.8	77.9	78.7	73.0	64.8
SWAHR [22]	CVPR 2021	HrHRNet-W48	640	71.6	88.5	77.6	78.9	72.4	63.0
CenterAttention [4]	ICCV 2021	HrHRNet-W48	640	67.6	87.7	72.7	73.9	68.2	60.3
Ours		HrHRNet-W48	640	**72.6**	88.8	78.9	79.2	73.1	65.6
w/ multi-scale testing									
HrHRNet [6]	CVPR 2020	HrHRNet-W48	640	67.6	87.4	72.6	75.8	68.1	58.9
DEKR [11]	CVPR 2021	HRNet-W48	640	68.0	85.5	73.4	76.6	68.8	58.4
PINet [37]	NIPS 2021	HRNet-W32	512	69.8	89.1	75.6	76.4	70.5	62.2
SWAHR [22]	CVPR 2021	HrHRNet-W48	640	73.8	90.5	79.9	81.2	74.7	64.7
CenterAttention [4]	ICCV 2021	HrHRNet-W48	640	69.4	88.6	74.6	76.6	70.0	61.5
Ours		HrHRNet-W48	640	**74.1**	90.7	80.2	81.3	74.9	65.1

images), COCO validation set (5k images), and COCO test-dev set (20k images). Following the train-test split of [22], we report results on the val2017 set and test-dev2017 set. Also following [22], we evaluate model performance using standard average precision (AP) calculated based on Object Keypoint Similarity (OKS) on this dataset, and report the following metrics: AP, AP^{50}, AP^{75}, AP^{M}, and AP^{L}.

Implementation details. Following [4, 22], we use the HrHRNet [6] as the baseline, and apply our proposed method to the respective two backbones including HrHRNet-W32 and HrHRNet-W48. For these backbones, we follow their original training and testing configurations specified in [6]. Also following [6], we adopt three scales 0.5, 1, and 2 in multi-scale testing. To calculate $\hat{L}_k$ following Eq. 21, we set the number of samples M to 256 and the hyperparameter U w.r.t. the finite range B_U to 64 in our experiments.

Results. In Tab. 1 and Tab. 2, we report single-scale testing and multi-scale testing results on the COCO val2017 set. In Tab. 3 and Tab. 4, we report single-scale testing and multi-scale testing results on the COCO test-dev2017 set. We observe that after applying our method on both HrHRNet-W32 and HrHRNet-W48, a significant performance improvement is achieved, demonstrating the effectiveness of our method. Moreover, we also compare our method with other state-of-the-art bottom-up human pose estimation methods. Compared to these methods, our method consistently achieves the highest AP score, further demonstrating the effectiveness of our method.

4.2. CrowdPose

Dataset & evaluation metric. The CrowdPose dataset [19] contains about 20k images and 80k person instances, which are annotated with 14 body joints. This dataset consists of three subsets including CrowdPose training set (10k images), CrowdPose validation set (2k images), and CrowdPose testing set (8k images). Following the train-test split of [6, 22], we report results on the testing set. Also following [6, 22], we evaluate model performance using standard AP calculated based on OKS on the CrowdPose dataset, and report the following metrics: AP, AP^{50}, AP^{75}, AP^{E}, AP^{M}, and AP^{H}.

Implementation details. On the CrowdPose dataset, we also use the HrHRNet [6] as the baseline, and we use HrHRNet-W48 as the backbone following [4,6,22]. We follow the original training and testing configurations specified in [6], and also follow [6] to adopt three scales 0.5, 1, and 2 in multi-scale testing. Besides, same as the experiments on the COCO dataset, we also set the number of samples M to 256 and the hyperparameter U w.r.t. the finite range B_U to 64 on the CrowdPose dataset.

Results. In Tab. 5, we report the single-scale testing and multi-scale testing results on the CrowdPose testing set. As shown, our method consistently achieves the highest AP score, demonstrating the effectiveness of our method.

4.3. Ablation Studies

We conduct ablation studies on the COCO validation set via applying our proposed method on HrHRNet-W32 [6] with single-scale testing.

Impact of the number of samples M. To calculate $\hat{L}_k$ following Eq. 21, we need to set the number of samples M, which we set to 256 in our experiments. We evaluate other choices of the number of samples M in Tab. 6. As shown, all variants outperform the baseline method, and after the number of samples M becomes larger than 256 the model performance becomes stabilized. Therefore, we set the number of samples M to be 256 in our experiments.

Impact of the finite range B_U with different U. We evaluate different choices of U in Tab. 7. As shown, all variants outperform the baseline method, and after the hyperparameter U becomes larger than 64, the model performance does

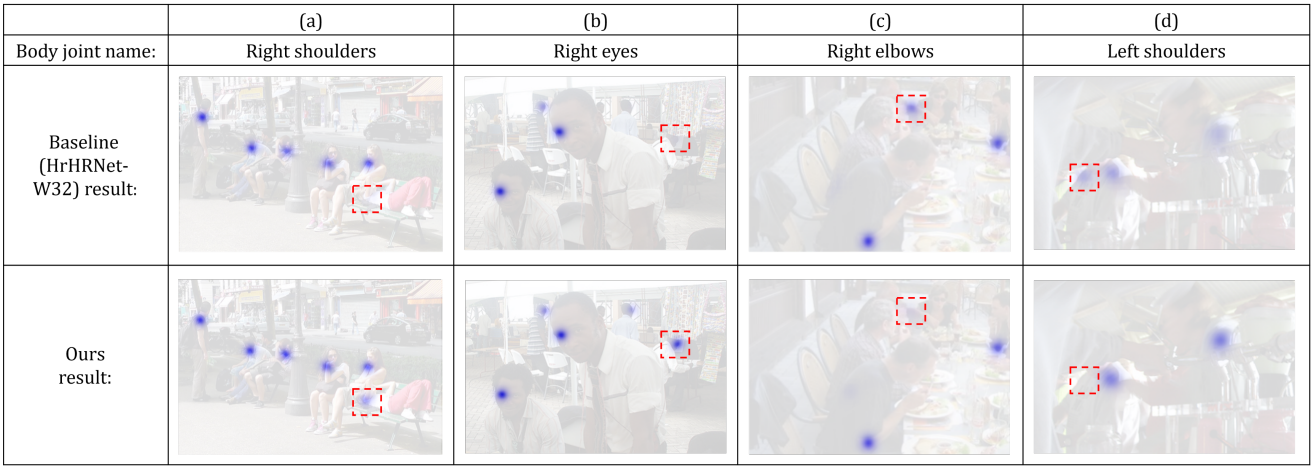

Figure 2. Qualitative results of our method and the baseline HrHRNet-W32 model [6]. As shown, the baseline method misses body joints (in (a) and (b)) or misidentifies body joints (in (c) and (d)) in some sub-regions of the predicted heatmap (see the sub-regions framed with dashed lines). Meanwhile, our method provides a more accurate localization result for the body joints of different people in different sub-regions of the predicted heatmap at the same time. More qualitative results are in the supplementary. (Best viewed in color.)

Table 6. Evaluation on the number of samples M.

Method	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
Baseline(HrHRNet-W32)	67.1	86.2	73.0	61.5	76.1
4 samples	67.9	86.9	73.8	62.4	76.9
16 samples	68.9	87.5	74.8	63.5	77.4
64 samples	69.6	87.9	75.6	63.9	77.8
256 samples	69.9	88.1	76.0	64.2	78.1
1024 samples	69.8	88.2	76.0	64.3	78.0

not enhance anymore. Thus, we set the hyperparameter U to be 64 in our experiments.

Table 7. Evaluation on the hyperparameter U w.r.t. the finite range B_U.

Method	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
Baseline(HrHRNet-W32)	67.1	86.2	73.0	61.5	76.1
$U = 8$	67.7	86.7	73.5	62.2	76.5
$U = 16$	68.6	87.3	74.2	62.9	76.9
$U = 32$	69.4	87.8	75.4	63.7	77.6
$U = 64$	69.9	88.1	76.0	64.2	78.1
$U = 128$	69.8	88.0	75.8	64.0	78.0

Training time. On the COCO dataset, we test the training time of our method that trains the backbone model (HrHRNet-W32 [6]) with the loss function in Eq. 22, and compare it with the training time of the baseline that trains the same network with the overall L2 loss. As shown in Tab. 8, though our method achieves much better performance, it brings only very little increase of the training time. Note that as we follow the same evaluation procedure of previous works [6, 22], the testing time with and without our proposed method are the same.

Qualitative results. Some qualitative results are shown in Fig. 2. As shown, the baseline method which uses the overall L2 loss to optimize the heatmap prediction can miss or get inaccurate body joints in some sub-regions of the predicted heatmap (see the sub-regions framed with dashed lines). In contrast, our method locates body joints of different people in different sub-regions of the predicted heatmap more accurately at the same time, demonstrating the effectiveness of our method.

Table 8. Comparison of the training time.

Method	Training time per epoch	Performance(AP)
Baseline(HrHRNet-W32)	1.11h	67.1
Baseline + Ours	1.19h	69.9

5. Conclusion

In this paper, we have proposed a novel bottom-up human pose estimation method that optimizes the heatmap prediction via minimizing the distance between two characteristic functions respectively constructed from the predicted and GT heatmaps. We theoretically analyze that the distance between the two characteristic functions is the upper bound of the L2 losses w.r.t. sub-regions of the predicted heatmap. Thus, via minimizing the distance between the two characteristic functions, our method locates body joints in different sub-regions of the predicted heatmap more accurately at the same time. Our method achieves superior performance on the COCO dataset and the CrowdPose dataset. Besides, our method could potentially also be applied in other tasks such as multi-object 6D pose estimation [1], facial landmark extraction [3], and fingerprint minutiae detection [10]. We leave this as our future work.

Acknowledgement. This work is supported by MOE AcRF Tier 2 (Proposal ID: T2EP20222-0035), National Research Foundation Singapore under its AI Singapore Programme (AISG-100E-2020-065), and SUTD SKI Project (SKI 2021_02_06).

References

[1] Arash Amini, Arul Selvam Periyasamy, and Sven Behnke. Yolopose: Transformer-based multi-object 6d pose estimation using keypoint regression. In *Intelligent Autonomous Systems 17: Proceedings of the 17th International Conference IAS-17*, pages 392–406. Springer, 2023. 8

[2] Abdul Fatir Ansari, Jonathan Scarlett, and Harold Soh. A characteristic function approach to deep implicit generative modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7478–7487, 2020. 3, 4, 5

[3] Matteo Bodini. A review of facial landmark extraction in 2d images and videos using deep learning. *Big Data and Cognitive Computing*, 3(1):14, 2019. 8

[4] Guillem Brasó, Nikita Kister, and Laura Leal-Taixé. The center of attention: Center-keypoint grouping via attention for multi-person pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11853–11863, 2021. 5, 6, 7

[5] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 1, 2, 3, 5, 6, 7

[6] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5386–5395, 2020. 1, 2, 5, 6, 7, 8

[7] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu. Skeleton-based action recognition with shift graph convolutional network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 183–192, 2020. 1

[8] Kacper P Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton. Fast two-sample testing with analytic representations of probability measures. *Advances in Neural Information Processing Systems*, 28, 2015. 3

[9] TW Epps and Kenneth J Singleton. An omnibus test for the two-sample problem using the empirical characteristic function. *Journal of Statistical Computation and Simulation*, 26(3-4):177–203, 1986. 3

[10] Yulin Feng and Ajay Kumar. Detecting locally, patching globally: An end-to-end framework for high speed and accurate detection of fingerprint minutiae. *IEEE Transactions on Information Forensics and Security*, 2023. 8

[11] Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang. Bottom-up human pose estimation via disentangled keypoint regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14676–14686, 2021. 3, 5, 6, 7

[12] Kerui Gu, Linlin Yang, and Angela Yao. Removing the bias of integral pose regression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11067–11076, 2021. 1

[13] CE Heathcote. A test of goodness of fit for symmetric random variables1. *Australian Journal of Statistics*, 14(2):172–181, 1972. 3

[14] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European conference on computer vision*, pages 34–50. Springer, 2016. 3

[15] Wentao Jiang, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, and Si Liu. Posetrans: A simple yet effective pose transformation augmentation for human pose estimation. *arXiv preprint arXiv:2208.07755*, 2022. 5, 6

[16] Sheng Jin, Wentao Liu, Enze Xie, Wenhai Wang, Chen Qian, Wanli Ouyang, and Ping Luo. Differentiable hierarchical graph grouping for multi-person pose estimation. In *European Conference on Computer Vision*, pages 718–734. Springer, 2020. 1, 2, 3, 5, 6

[17] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11977–11986, 2019. 1, 2, 3, 5, 6

[18] Jia Li, Wen Su, and Zengfu Wang. Simple pose: Rethinking and improving a bottom-up approach for multi-person pose estimation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11354–11361, 2020. 6

[19] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10863–10872, 2019. 2, 6, 7

[20] Shengxi Li, Zeyang Yu, Min Xiang, and Danilo Mandic. Reciprocal adversarial learning via characteristic functions. *Advances in Neural Information Processing Systems*, 33:217–228, 2020. 3

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 6

[22] Zhengxiong Luo, Zhicheng Wang, Yan Huang, Liang Wang, Tieniu Tan, and Erjin Zhou. Rethinking the heatmap regression for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13264–13273, 2021. 1, 2, 5, 6, 7, 8

[23] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. *Advances in neural information processing systems*, 30, 2017. 1, 2, 3, 6

[24] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 1, 2, 6

[25] Xuecheng Nie, Jiashi Feng, Jianfeng Zhang, and Shuicheng Yan. Single-stage multi-person pose machines. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6951–6960, 2019. 1, 2, 3, 6

[26] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. 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)*, pages 269–286, 2018. 1, 2, 3, 5, 6

[27] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4929–4937, 2016. 3

[28] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. Pose-normalized image generation for person re-identification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 650–667, 2018. 1

[29] Haoxuan Qu, Li Xu, Yujun Cai, Lin Geng Foo, and Jun Liu. Heatmap distribution matching for human pose estimation. In *Advances in Neural Information Processing Systems*. 2

[30] Dahu Shi, Xing Wei, Liangqi Li, Ye Ren, and Wenming Tan. End-to-end multi-person pose estimation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11069–11078, 2022. 5, 6, 7

[31] Weibo Shu, Jia Wan, Kay Chen Tan, Sam Kwong, and Antoni B Chan. Crowd counting in the frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19618–19627, 2022. 3

[32] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019. 1, 2

[33] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. *Advances in neural information processing systems*, 27, 2014. 1, 2

[34] Ali Varamesh and Tinne Tuytelaars. Mixture dense regression for object detection and human pose estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13083–13092. IEEE, 2020. 6

[35] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. Pose-aware multi-level feature network for human object interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9469–9478, 2019. 1

[36] Dongkai Wang and Shiliang Zhang. Contextual instance decoupling for robust multi-person pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11060–11068, 2022. 5, 6, 7

[37] Dongkai Wang, Shiliang Zhang, and Gang Hua. Robust pose estimation in crowded scenes with direct pose-level inference. *Advances in Neural Information Processing Systems*, 34:6278–6289, 2021. 5, 6, 7

[38] Fangyun Wei, Xiao Sun, Hongyang Li, Jingdong Wang, and Stephen Lin. Point-set anchors for object detection, instance segmentation and pose estimation. In *European Conference on Computer Vision*, pages 527–544. Springer, 2020. 5, 6

[39] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018. 1, 2

[40] Yabo Xiao, Dongdong Yu, Xiao Juan Wang, Lei Jin, Guoli Wang, and Qian Zhang. Learning quality-aware representation for multi-person pose regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2822–2830, 2022. 5, 6

[41] Jiangtao Xie, Fei Long, Jiaming Lv, Qilong Wang, and Peihua Li. Joint distribution matters: Deep brownian distance covariance for few-shot classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7972–7981, 2022. 3

[42] Nan Xue, Tianfu Wu, Gui-Song Xia, and Liangpei Zhang. Learning local-global contextual adaptation for multi-person pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13065–13074, 2022. 5, 6

[43] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*, 2018. 1

[44] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution transformer for dense prediction. 2021. 1, 2

[45] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 3