SimCC: a Simple Coordinate Classification Perspective for Human Pose Estimation

Yanjie Li[1] Sen Yang[2] Peidong Liu[1] Shoukui Zhang[3] Yunxiao Wang[1]
Zhicheng Wang[4] Wankou Yang[2] Shu-Tao Xia[1,5⋆]

[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]Southeast University [3]Meituan Inc. [4]Nreal
[5]Research Center of Artificial Intelligence, Peng Cheng Laboratory

Abstract. The 2D heatmap-based approaches have dominated Human Pose Estimation (HPE) for years due to high performance. However, the long-standing quantization error problem in the 2D heatmap-based methods leads to several well-known drawbacks: 1) The performance for the low-resolution inputs is limited; 2) To improve the feature map resolution for higher localization precision, multiple costly upsampling layers are required; 3) Extra post-processing is adopted to reduce the quantization error. To address these issues, we aim to explore a brand new scheme, called *SimCC*, which reformulates HPE as two classification tasks for horizontal and vertical coordinates. The proposed SimCC uniformly divides each pixel into several bins, thus achieving *sub-pixel* localization precision and low quantization error. Benefiting from that, SimCC can omit additional refinement post-processing and exclude upsampling layers under certain settings, resulting in a more simple and effective pipeline for HPE. Extensive experiments conducted over COCO, CrowdPose, and MPII datasets show that SimCC outperforms heatmap-based counterparts, especially in low-resolution settings by a large margin. Code is now publicly available at https://github.com/leeyegy/SimCC.

Keywords: Human Poes Estimation, 2D Heatmap, Coordinate Classification

1 Introduction

2D Human Pose Estimation (HPE) aims to localize body joints from a single image, where 2D heatmap-based methods [2, 3, 6, 7, 43, 17, 18, 20, 23, 29, 38, 40] has become the *de facto* standard in recent years. The 2D heatmap is generated as a 2-dimensional Gaussian distribution centering at the ground-truth joint position, which inhibits the cases of false positive and smooths the training process by assigning a probability value to each position.

Despite its success, heatmap-based methods suffer seriously from the long-standing quantization error problem, which is caused by mapping the continuous coordinate values into discretized 2D downscaled heatmaps. The substantial quantization error further brings about several well-known shortcomings: 1)

⋆ Corresponding author: Shu-Tao Xia (xiast@sz.tsinghua.edu.cn).

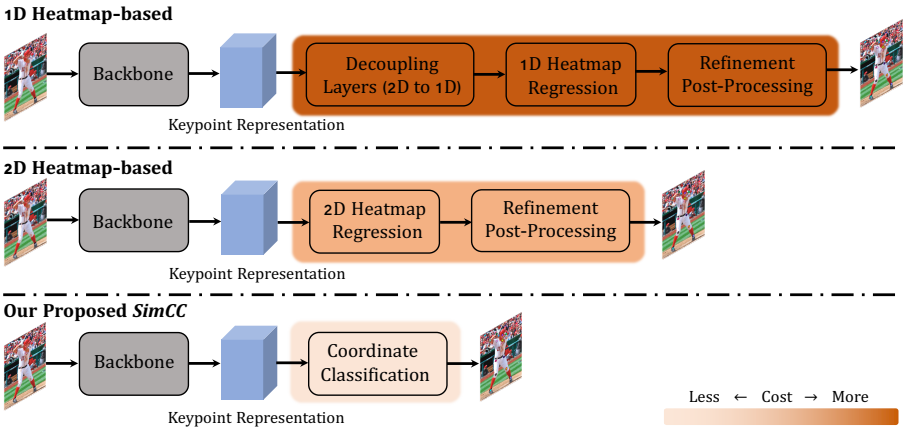


Fig. 1. Comparisons between the proposed SimCC and 2D/1D heatmap-based pipelines. The 2D heatmap-based scheme includes: 1) a backbone to extract keypoint representations; 2) a regression head to generate the 2D heatmap, which may consist of multiple time-consuming upsampling layers; 3) extra post-processing to refine the predictions, such as empirical shift and DARK [43]. The 1D heatmap regression [42] is introduced for facial landmark. Compared to the 2D heatmap-based scheme, The 1D heatmap regression [42] brings additional learnable decoupling layers consisting of multiple CNN layers and a co-attention module, to transform the 2D features to 1D heatmaps. Different from these heatmap-based schemes, the proposed SimCC is much simpler, which only needs two linear classifier heads for coordinate classification.

Costly upsampling layers (*e.g.*, deconvolution layers [38]) are used to increase the feature map resolution to alleviate the quantization error; 2) Extra post-processing is introduced to refine the predictions; 3) The performances are far from satisfactory for low-resolution inputs due to the serious quantization error. Considering obtaining high-resolution 2D heatmap brings heavy computation cost, a natural way to decrease the quantization error is to firstly decouple the 2D heatmap into 1D heatmap and then increase the resolution, which has been explored by Yin et al. [42] for facial landmark area. However, to realize that goal, Yin et al. [42] introduces additional decoupling layers and costly deconvolution modules, resulting in an even more complicated pipeline.

Therefore, in this work, we try to explore a brand-new scheme against heatmap-based methods for HPE. Specifically, we propose a simple yet effective coordinate classification pipeline, namely *SimCC*, which regards HPE as two classification tasks for horizontal and vertical coordinates. SimCC firstly employs a Convolutional Neural Network (CNN) or Transformer-based backbone to extract keypoint representations. Given the obtained keypoint representations, SimCC then performs coordinate classification for vertical and horizontal coordinates independently to yield the final predictions. To reduce the quantization error, SimCC uniformly divides each pixel into several bins, which achieves *sub-pixel* localization precision. Note that different from heatmap-based approaches which

may introduce multiple deconvolution layers, SimCC only needs two lightweight classifier heads (*i.e.* only one linear layer for each head).

Fig. 1 shows the comparisons between our proposed SimCC and 1D/2D heatmap-based approaches. Compared to the dominant 2D heatmap-based scheme, SimCC has three benefits: 1) It reduces quantization error by uniformly dividing each pixel into several bins; 2) SimCC omits upsampling layers under certain settings [38] and excludes the costly refinement post-processing, which is more friendly to real-world applications; 3) SimCC shows impressing performance even with low input sizes. Our contributions are summarized as follows:

- We propose a coordinate classification pipeline for human pose estimation called SimCC, reformulating the problem as two classification tasks for horizontal and vertical coordinates. SimCC serves as a general scheme and can be easily applied to existing CNN-based or Transformer-based HPE models.
- SimCC achieves high efficiency by omitting the extra time-consuming upsampling and post-processing in heatmap-based methods. In particular, applying SimCC reduces over 55% GFLOPs of SimBa-Res50 [38] and achieves higher model performance than heatmap-based counterpart.
- Comprehensive experiments over COCO, CrowdPose, and MPII datasets are conducted to verify the effectiveness of the proposed SimCC with different backbones and multiple input sizes.

It's our belief that the predominant 2D heatmap-based methods may not be the final solution for HPE due to its high computation cost, complicated post-processing and poor performance under low input resolutions. We hope that the exploration of SimCC could provide a new perspective for the potential research work and practical deployment for HPE.

2 Related Work

Regression-based HPE. Regression-based methods [35, 4, 33, 31, 30, 24] are explored more often in the early stage of 2D human pose estimation. Different from relying on 2D grid-like heatmap, this line of work directly regresses the keypoint coordinates in a computationally friendly framework. However, only a handful of existing methods adopt this scheme due to the unsatisfactory performance. Very recently, Li *et al.* [14] introduce the residual log-likelihood (RLE), which utilizes the normalizing flows [27] to capture the underlying output distribution and makes regression-based methods match the accuracy of state-of-the-art heatmap-based methods. Our method focuses on the coordinate representation, while the core idea of RLE is to construct an adaptive loss based on the normalizing flows, which is complementary to our work.

2D heatmap-based HPE. Another line of work [2, 3, 6, 7, 17, 18, 20, 23, 29, 38, 40, 43, 11] adopts two-dimensional Gaussian distribution (*i.e.*, *heatmap*) to represent joint coordinate. Each position on the heatmap is assigned with a probability to be the ground truth point. As one of the earliest uses of heatmap, Tompson *et al.* [34] propose a hybrid architecture consisting of a deep Convolutional Network and a Markov Random Field. Newell *et al.* [23] introduce hourglass-style archi-

tecture into HPE. Papandreou *et al.* [25] propose to aggregate the heatmap and offset prediction to improve the localization precision. Xiao *et al.* [38] propose a simple baseline that utilizes three deconvolutional layers following a backbone network to obtain the final predicted heatmap. Instead, Sun *et al.* [29] propose a novel network to maintain high-resolution representations through the whole process, achieving significant improvement. Owing to the involvement of spatial uncertainty, this kind of learning schema has the tolerance of mistakes of jitter. As a result, heatmap-based methods keep stable state-of-the-art performance for years. However, quantization error remains a significant problem of the heatmap-based methods, especially in low input resolutions. To address the large quantization error caused by the discretized 2D downscaled heatmaps, Zhang *et al.* [43] propose to comprehensively account for the distribution information of heatmap activation by adopting Taylor-expansion based distribution approximation as post-processing, which complicates the pipeline.

1D heatmap regression in facial landmark. Outside the realm of human pose estimation, 1D heatmap-based methods [42, 39] have been explored for facial landmark detection. Among those, Yin *et al.* [42] propose an attentive 1D heatmap regression method, which adopts learnable decoupling layers to transform 2D heatmap to 1D heatmap and then uses additional deconvolution layers to alleviate the quantization error. To capture the joint distribution information between the decoupled 1D heatmaps, a co-attention module is introduced in the 1D heatmap regression [42].

Coordinate classification. Concurrent to our work, Chen *et al.* [5] propose Pix2Seq to casts object detection as a language modeling task, where an object is described as sequences of five discrete tokens for further classification. In Pix2Seq, the Transformer decoder architecture is essential to "read out" each object (yield the predictions). By contrast, our proposed SimCC aims to explore a new path against heatmap-based methods for human pose estimation, which can be easily combined with CNN or Transformer-based HPE methods and does not rely on an additional Transformer decoder to generate the prediction.

3 SimCC: Reformulating HPE from Classification Perspective

The key idea of SimCC is to regards human pose estimation as two classification tasks for vertical and horizontal coordinates and to reduce quantization error by dividing each pixel to multiple bins. Fig. 2 shows the schematic illustration of SimCC composed of a backbone network and two classifier heads. We will describe each components in this section in details.

Backbone. Given an input image of size $H \times W \times 3$, SimCC employs either CNN-based or Transformer-based network (*e.g.*, HRNet [29], TokenPose [18]) as the backbone to extract n keypoint representations for n corresponding keypoints.

Head. As shown in Fig. 2, horizontal and vertical classifiers (*i.e.*, only one linear layer for each classifier) are appended after the backbone to perform coordinate classification, respectively. For the CNN-based backbone, we simply flatten the

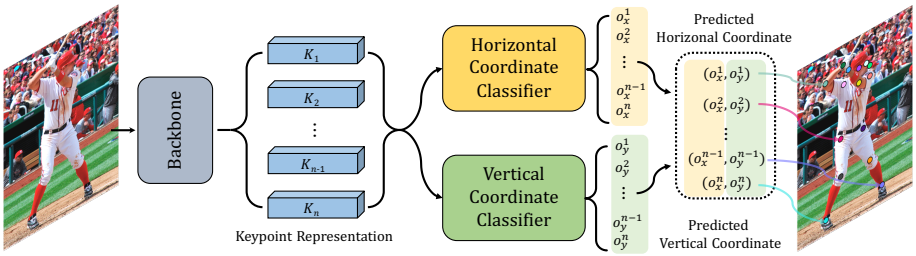
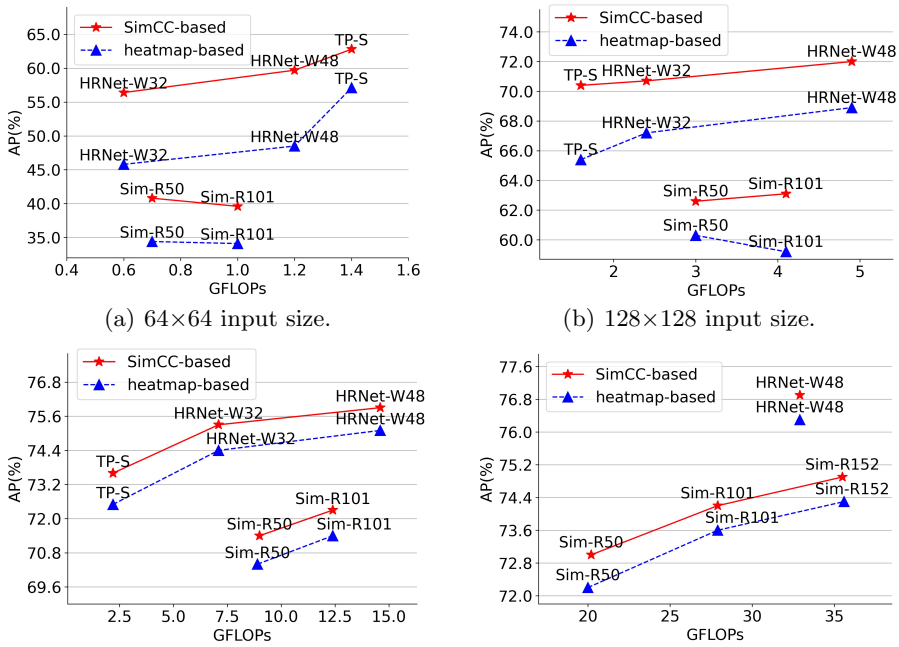


Fig. 2. The proposed SimCC pipeline for HPE. SimCC firstly extracts n keypoint representations via a backbone which can be either CNN-based or Transformer-based networks. For the CNN-based backbone, we simply flatten the obtained keypoint representations from (n, H', W') to $(n, H' \times W')$ for the subsequent classification. Based on the n keypoint representations, SimCC then performs coordinate classification for horizontal and vertical axes independently to yield the final predictions. Specifically, given i-th keypoint representation as input, the horizontal and vertical coordinate classifiers (*i.e.*, only one linear layer for each classifier) generate the i-th keypoint predictions o_x^i and o_y^i, respectively. Note that SimCC uniformly divides each pixel into multiple bins thus the quantization error is reduced and sub-pixel localization precision is achieved.

outputted keypoint representations from (n, H', W') to $(n, H' \times W')$ for classification. Compared to heatmap-based approach [38] which uses multiple costly deconvolution layers as head, SimCC head is much more lightweight and simple.

Coordinate classification. To achieve classification, we propose to uniformly discretize each continuous coordinate value into an integer as class label for model training: $c_x \in [1, N_x], c_y \in [1, N_y]$, where $N_x = W \cdot k$ and $N_y = H \cdot k$ represent the number of bins for horizontal and vertical axes, respectively. k is the splitting factor and set as ≥ 1 to reduce quantization error, resulting in *sub-pixel* localization precision. To yield the final prediction, SimCC performs vertical and horizontal coordinate classification independently based on the n keypoint representations learnt by the backbone. Concretely, given i-th keypoint representation as input, the i-th keypoint predictions o_x^i and o_y^i are generated by the horizontal and vertical coordinate classifiers, respectively. In addition, Kullback–Leibler divergence is used as loss function for training.

Label smoothing. In traditional classification tasks, label smoothing [32] is widely utilized to enhance model performance. Hence, we adopt it for SimCC, which is called *equal label smoothing* in this paper. However, equal label smoothing punishes the false labels indiscriminately, which has ignored the spatial relevance of adjacent labels for the task of human pose estimation. A more reasonable solution is supposed to encourage the model to work in this way: the closer the output category is to the groundtruth, the better. To address this issue, we also explore to use *Laplace* or *Gaussian label smoothing*, resulting in smoothed labels following corresponding distribution. Unless noted otherwise, SimCC is used as the abbreviation for the variant with equal label smoothing .



(a) 64×64 input size. (b) 128×128 input size.

(c) 256×192 input size. (d) 384×288 input size.

Fig. 3. Comparisons with 2D heatmap-based approaches on COCO2017 val set across various input sizes. 'TP-S' and 'Sim' represent TokenPose-S [18] and SimpleBaseline [38], respectively. Both CNN-based (*i.e.*, SimpleBaseline [38] and HRNet [29]) and Transformer-based (*i.e.*, TokenPose [18]) HPE models are chosen to verify the effectiveness of the proposed SimCC. SimCC shows clear gains compared to 2D heatmap-based counterparts across various input sizes, especially in low resolution.

3.1 Comparisons to 2D heatmap-based approaches

In this part, we give a comprehensive investigation on the superiority of using SimCC scheme compared to the 2D heatmap-based approaches.

Quantization error. Due to the computational cost of obtaining or maintaining high-resolution two-dimensional structure, 2D heatmap-based methods tend to output feature maps with $\lambda\times$ downscaled input resolution, which significantly enlarges the quantization error. On the contrary, SimCC uniformly divides each pixel into k (≥ 1) bins during discretization, which reduces the quantization error and obtains sub-pixel localization precision.

Refinement post-processing. Heatmap-based approaches rely heavily on extra post-processing (*e.g.*, empirical shift and DARK [43]) to reduce the quantization error. As shown in Table 1, the performance of heatmap-based methods drops significantly if without using post-processing for refinement. However, these post-processing strategies are usually computationally expensive and thus unfriendly to real-world applications. For example, DARK [43] uses Taylor-

expansion and higher derivative needs to be calculated based on the obtained 2D heatmap. By contrast, the proposed SimCC omits refinement post-processing due to its sub-pixel localization precision, resulting in a simpler scheme for HPE.
Low/high resolution robustness. Fig. 3 visualizes the comparison results. Benefiting from low quantization error, SimCC-based methods can easily outperform heatmap-based counterparts in various input sizes (*i.e.*, 64×64, 128×128, 256×192 and 384×288), demonstrating clear gains especially in low input resolutions. More specific quantitative results are discussed in Section 4.
Speed. SimCC makes methods like [38] get rid of time-consuming deconvolution modules, which can speed up the inference. It's worth noting that after removing the upsampling layers, SimpleBaseline-Res50 [38] with SimCC reduces **57.3%** GFLOPs, improves **23.5%** speed, and gains **+0.4** AP over heatmap-based counterpart. More comparisons are presented in Section 4.

4 Experiments

In the following sections, we empirically investigate the effectiveness of the proposed SimCC for 2D human pose estimation. We conduct experiments on three benchmark datasets: COCO [19], CrowdPose [15], and MPII [1].

4.1 COCO Keypoint Detection

As one of the largest and most challenging datasets for HPE, the COCO dataset [19] contains more than 200,000 images and 250,000 person instances labeling with 17 keypoints (*e.g.*, nose, left ear, etc.). The COCO dataset is divided into three parts: 57k images for the training set, 5k for val set and 20k for test-dev set. In this paper, we follow the data augmentation in [29].
Evaluation metric. The standard average precision (AP) is used as our evaluation metric on the COCO dataset, which is calculated based on Object Keypoint Similarity (OKS):

$$OKS = \frac{\sum_i exp(-d_i^2/2s^2j_i^2)\sigma(v_i > 0)}{\sum_i \sigma(v_i > 0)}, \tag{1}$$

where d_i is the Euclidean distance between the i-th predicted keypoint coordinate and its corresponding coordinate groundtruth, j_i is a constant, v_i is the visibilty flag, σ denotes indicator function and s is the object scale.
Baselines. There are many *CNN-based* and recent *Transformer-based* methods for HPE. To show the superiority of the proposed SimCC, we choose two state-of-the-art methods (*i.e.*, SimpleBaseline [38] and HRNet [29]) from the former and one (*i.e.*, TokenPose [18]) from the latter as our baselines.
Implementation details. For the selected baselines, we simply follow the original settings in their papers. Specifically, for SimpleBaseline [38], the base learning rate is set as $1e-3$, and is dropped to $1e-4$ and $1e-5$ at the 90-th and 120-th epochs respectively. For HRNet [29], the base learning rate is set as

Table 1. Comparisons with 2D heatmap-based methods on COCO validation set. The results are provided with the same detected human boxes for fair comparison. 2D heatmap-based approaches adopt extra post-processing for refinement following their original paper, *i.e.*, DARK [43] for TokenPose [18] and empirical shift for HRNet [29] as well as SimpleBaseline [38]. SimCC brings significant gains in all input resolutions while omitting costly refinement post-processing. In particular, SimCC-based SimBa-R50 [38] achieves better results than 2D heatmap-based counterpart with over **55%** FLOPs reduction.

Method	Scheme	Input size	#Params	GFLOPs	Extra post.	AP	AR
TokenPose-S [18]	Heatmap	64×64	4.9M	1.4	w/o	35.9	47.0
	Heatmap	64×64	4.9M	1.4	DARK [43]	57.1	64.8
	SimCC	64×64	4.9M	1.4	**w/o**	**62.8**	**70.1**
	Heatmap	128×128	5.2M	1.6	w/o	57.6	64.9
	Heatmap	128×128	5.2M	1.6	DARK [43]	65.4	71.6
	SimCC	128×128	5.1M	1.6	**w/o**	**70.4**	**76.4**
	Heatmap	256×192	6.6M	2.2	w/o	69.9	75.8
	Heatmap	256×192	6.6M	2.2	DARK [43]	72.5	78.0
	SimCC	256×192	5.5M	2.2	**w/o**	**73.6**	**78.9**
SimBa-R50 [38]	Heatmap	64×64	34.0M	0.7	w/o	25.8	36.0
	Heatmap	64×64	34.0M	0.7	shift	34.4	43.7
	SimCC	64×64	24.7M	0.3	**w/o**	**39.3**	**48.4**
	Heatmap	128×128	34.0M	3.0	w/o	55.4	63.3
	Heatmap	128×128	34.0M	3.0	shift	60.3	67.6
	SimCC	128×128	25.0M	1.3	**w/o**	**62.6**	**69.5**
	Heatmap	256×192	34.0M	8.9	w/o	68.5	74.8
	Heatmap	256×192	34.0M	8.9	shift	70.4	76.3
	SimCC	256×192	25.7M	3.8	**w/o**	**70.8**	**76.8**
HRNet-W48 [29]	Heatmap	64×64	63.6M	1.2	w/o	36.9	47.8
	Heatmap	64×64	63.6M	1.2	shift	48.5	57.8
	SimCC	64×64	63.7M	1.2	**w/o**	**59.7**	**67.5**
	Heatmap	128×128	63.6M	4.9	w/o	63.3	70.5
	Heatmap	128×128	63.6M	4.9	shift	68.9	75.3
	SimCC	128×128	64.1M	4.9	**w/o**	**72.0**	**77.9**
	Heatmap	256×192	63.6M	14.6	w/o	73.1	78.7
	Heatmap	256×192	63.6M	14.6	shift	75.1	80.4
	SimCC	256×192	66.3M	14.6	**w/o**	**75.9**	**81.2**

$1e-3$, and decreased to $1e-4$ and $1e-5$ at the 170-th and 200-th epochs. The total training processes are terminated within 140 and 210 epochs respectively for SimpleBaseline [38] and HRNet [29]. Note that the training process of TokenPose-S follows [29]. In this paper, we use the two-stage [29, 38, 6, 25] top-down human pose estimation pipeline: the person instances are firstly detected and then the keypoints are estimated. We adopt a popular person detector with 56.4% AP provided by [38] for COCO validation set. In addition, label smoothing is adopted in model training, which is commonly used in the task of classification

for better generalization (equal smoothing sets the coefficient as 0.1 by default, following [32]). Experiments are conducted in 4 NVIDIA Tesla V100 GPUs.

Table 2. Comparisons on the COCO test-dev set. 'Trans.' represents Transformer [36] for short. "†" indicates that the Gaussian label smoothing is adopted. The proposed SimCC achieves state-of-the-art results, demonstrating clear performance improvements compared to 2D heatmap-based counterparts.

Method	Encoder	Input size	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
2D Heatmap-based									
Mask-RCNN [9]	ResNet-50-FPN	-	-	63.1	87.3	68.7	57.8	71.4	-
CMU-Pose [3]	VGG-19 [28]	-	-	64.2	86.2	70.1	61.0	68.8	-
G-RMI [25]	ResNet-101 [10]	353×257	-	64.9	85.5	71.3	62.3	70.0	69.7
AE [22]	Hourglass [23]	512×512	-	65.5	86.8	72.3	60.6	72.6	70.2
MultiPoseNet [13]	-	480×480	-	69.6	86.3	76.6	65.0	76.3	73.5
RMPE [8]	PyraNet [41]	320×256	26.7	72.3	89.2	79.1	68.0	78.6	-
CPN [6]	ResNet-Inception	384×288	29.2	72.1	91.4	80.0	68.7	77.2	78.5
CFN [12]	-	-	-	72.6	86.1	69.7	78.3	64.1	-
SimBa [38]	ResNet-152	384×288	35.6	73.7	91.9	81.1	70.3	80.0	79.0
TransPose-H [40]	HRNet-W48+Trans.	256×192	21.8	75.0	92.2	82.3	71.3	81.1	80.1
HRNet-W32 [29]	HRNet-W32	384×288	16.0	74.9	92.5	82.8	71.3	80.9	80.1
SimBa [38]	ResNet-50	384×288	20.0	71.5	91.1	78.7	67.8	78.0	76.9
HRNet-W48 [29]	HRNet-W48	256×192	14.6	74.2	92.4	82.4	70.9	79.7	79.5
HRNet-W48 [29]	HRNet-W48	384×288	32.9	75.5	92.5	83.3	71.9	81.5	80.5
Regression-based									
SPM [24]	Hourglass [23]	-	-	66.9	88.5	72.9	62.6	73.1	-
DeepPose [35]	ResNet-101	256×192	7.7	57.4	86.5	64.2	55.0	62.8	-
DeepPose [35]	ResNet-152	256×192	11.3	59.3	87.6	66.7	56.8	64.9	-
CenterNet [44]	Hourglass [23]	-	-	63.0	86.8	69.6	58.9	70.4	-
DirectPose [33]	ResNet-50	-	-	62.2	86.4	68.2	56.7	69.8	-
PointSetNet [37]	HRNet-W48	-	-	68.7	89.9	76.3	64.8	75.3	-
Integral Pose [31]	ResNet-101	256×256	11.0	67.8	88.2	74.8	63.9	74.0	-
TFPose [21]	ResNet-50+Trans.	384×288	20.4	72.2	90.9	80.1	69.1	78.8	-
PRTR [16]	HRNet-W48+Trans.	-	-	64.9	87.0	71.7	60.2	72.5	74.1
PRTR [16]	HRNet-W32+Trans.	384×288	21.6	71.7	90.6	79.6	67.6	78.4	78.8
PRTR [16]	HRNet-W32+Trans.	512×384	37.8	72.1	90.4	79.6	68.1	79.0	79.4
RLE [14]	HRNet-W48	-	-	75.7	92.3	82.9	72.3	81.3	-
SimCC-based									
SimBa (SimCC†)	ResNet-50	384×288	20.2	72.7	91.2	80.1	69.2	79.0	78.0
HRNet (SimCC†)	HRNet-W48	256×192	14.6	75.4	92.4	82.7	71.9	81.3	80.5
HRNet (SimCC†)	HRNet-W48	384×288	32.9	**76.0**	**92.4**	**83.5**	**72.5**	**81.9**	**81.1**

Results on the COCO val set. Extensive experiments are conducted on the COCO2017 validation set for comparing 2D heatmap-based and SimCC-based methods across various input resolutions (*i.e.*, 64×64, 128×128, 256×192, and 384×288). Note that the evaluation and network training are under the same input size. Some top-performed CNN-based and Transformer-based methods are chosen as our baselines. Results presented in Table 1 demonstrate that SimCC-based methods show consistent performance superiority over heatmap-based

counterparts, especially in low-resolution input cases. For example, SimCC-based HRNet-W48 [29] outperforms heatmap-based counterpart by **+0.8** AP at the input size of 256×192. And under low input resolution as 64×64, our SimCC shows much larger performance gain, *i.e.*, **+11.2** AP on the COCO val dataset.

According to the results presented in Table 1, we can further draw the following conclusions: 1) Heatmap-based approaches rely seriously on post-processing for refinement, which brings extra computational cost and complicates the whole process. For example, TokenPose-S dramatically drops **21.2** AP at the input size of 64×64 if without the DARK [43] post-processing; 2) Our proposed SimCC works well without any refinement post-processing, leading to a more simple and efficient scheme compared to heatmap-based methods. For instance, our SimCC-based HRNet-W48 w/o extra post-processing outperforms heatmap-based counterpart (empirical shift is used) by 0.8 AP at the input size of 256×192.

Results on the COCO test-dev set. We conduct comparisons on COCO test-dev set and present the results in Table 2. SimCC-based HRNet-W48 and SimpleBaseline-Res50 surpass heatmap-based counterparts by **+0.5** and **+1.2** AP respectively, at the input size of 384×288.

Inference speed. We discuss the impact of our proposed SimCC to the inference speed for SimpleBaseline [38], TokenPose-S [18] and HRNet-W48 [29]. The 'inference speed' here refers to the average time consuming of model feedforward (we compute 300 samples with batchsize=1). We adopt FPS to quantitatively illustrate the inference latency. The CPU implementation results are presented with the same machine: Intel(R) Xeon(R) Gold 6130 CPU @ 2.10GHz.

1) *SimpleBaseline* Adopting SimCC allows one to remove the costly deconvolution layers of SimpleBaseline. We conduct experiments via SimpleBaseline-Res50 [38] on COCO val set with input size of 256×192. SimCC-based method w/o deconvolution modules can surpass 2D heatmap-based counterpart by **+0.4** AP (70.8 vs. 70.4) with **23.5%** faster speed (21 vs. 17 FPS). More specific ablation study of deconvolution modules is conducted in Section 4.2.

2) *TokenPose&HRNet* Due to that SimpleBaseline [38] uses an encoder-decoder architecture, we can replace its decoder part (deconvolutions) with classifier heads of SimCC. But for HRNet [29] and TokenPose [18], they have no extra independent modules as the decoder. To apply SimCC to them, we directly append the classifier heads to the original HRNet and replace the MLP head of TokenPose with ours, respectively. These are minor changes to the original architectures, thus only bringing little computation overhead for HRNet [29] and even reducing the model parameters for TokenPose [18], as shown in Table 1. Hence, SimCC only has a slight impact on the inference latency for HRNet and TokenPose. For instance, the FPS of HRNet-W48 using heatmap or SimCC is almost the same (4.5/4.8) at the input size of 256×192.

Is 1D heatmap regression a promising solution for HPE? We also study the performance of expanding the 1D heatmap [42] into the field of HPE, which is initially designed for facial landmark. Table 3 shows that the 1D heatmap regression [42] increases the model parameters and computational cost yet performs even worse than 2D heatmap-based counterpart. The potential reason

Table 3. Comparisons with the 1D heatmap regression [42] and 2D heatmap-based methods. Results achieved by SimBa-R50 [38] via different schemes (2D/1D heatmap, SimCC) on COCO val set with input size of 256×256. SimCC performs better than 2D/1D heatmap-based methods and requires **only 41.7%/34.7%** computation cost.

Scheme	Deconvolution	#Params	GFLOPs	Extra post.	AP
2D Heatmap	3	34.0M	12.0	shift	70.4
2D Heatmap	3	34.0M	12.0	w/o	68.8
1D Heatmap [42]	5	39.0M	14.4	w/o	68.5
SimCC	**0**	**26.3M**	**5.0**	w/o	**70.4**

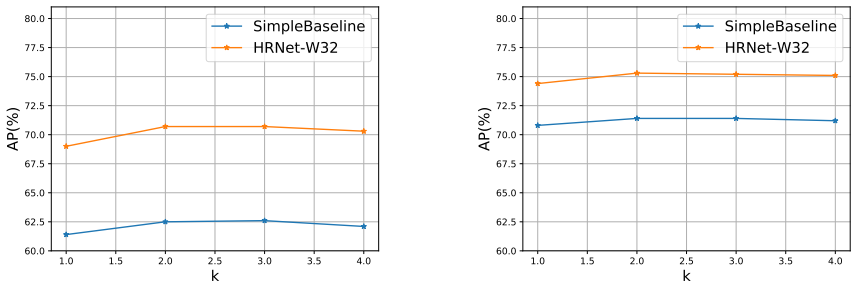
might be that the core challenges of facial landmark and HPE are different: facial landmark possesses rigid deformation while human body joints have much higher degrees of freedom. Since the co-attention module as well as decoupling layers in [42] are only empirically verified to be effective for the task of facial landmark and their generalization to other fields (*e.g.*, HPE) remains unclear.

4.2 Ablation Study

Splitting factor k**.** The splitting factor k controls the how many bins per pixel in SimCC. Specifically, the larger k is, the smaller the quantization error of SimCC is. Nevertheless, model training becomes more difficult when k increases. Hence, there is a trade-off between the quantization error and the model performance. We test $k \in \{1, 2, 3, 4\}$ based on SimpleBaseline [38] and HRNet [29] under various input resolutions. As shown in Fig. 4, model performance tends to increase first and then decrease as k grows. For HRNet-W32 [29], the recommended settings are $k = 2$ for both 128×128 and 256×192 input size. For SimBa-Res50 [38], the recommended settings are $k = 3$ and $k = 2$ for 128×128 and 256×192 input size, respectively.

Upsampling modules. Upsampling modules are usually computationally costly and substantially slow down the network's inference speed, however, indispensable for heatmap-based methods. Hence, it is of practical significance to explore if applying SimCC can reduce the dependence of upsampling modules in HPE. Notice that the upsampling modules[1] adopted in SimpleBaseline [38] is independent to the backbone and thus can be easily removed. Therefore we conduct ablation study of SimCC w/ and w/o deconvolution modules based on SimpleBaseline [38]. Table 4 show the results on the COCO2017 val dataset. It can be observed that compared to heatmap, SimCC allows one to remove the costly deconvolution layers of SimpleBaseline, resulting in consistent computational cost reduction across various input resolutions. For example, SimCC-based SimpleBaseline-Res50 w/o upsampling modules still outperform heatmap-based

[1] The upsampling modules used in SimpleBaseline [38] recover the feature map resolution from 1/32× to 1/4× input size, consisting of three deconvolution layers.



(a) 128×128 input size. (b) 256×192 input size.

Fig. 4. Ablation study of splitting factor k value on the COCO validation set. SimpleBaseline [38] uses ResNet-50 as backbone. k controls how many bins per pixel in SimCC. Model performance increases first and then drops as k becomes larger.

Table 4. Ablation study of upsampling modules. Results achieved by SimBa-R50 [38] on COCO val set. "Heatmap" represents 2D heatmap-based methods for short. Employing deconvolution improves SimCC-based methods yet the gains are slight. Even without any deconvolution layers, SimCC-based approaches surpass 2D heatmap-based counterparts, significantly reducing over 55% FLOPs.

Scheme	Input size	Deconvolution	#Params	GFLOPs	AP
Heatmap	64×64	✓	34.0M	0.7	34.4
SimCC	64×64	✓	34.1M	0.7	**40.8**
SimCC	64×64	✗	**24.7M**	**0.3**	39.3
Heatmap	128×128	✓	34.0M	3.0	60.3
SimCC	128×128	✓	34.8M	3.0	**62.6**
SimCC	128×128	✗	**25.0M**	**1.3**	**62.6**
Heatmap	256×192	✓	34.0M	8.9	70.4
SimCC	256×192	✓	36.8M	9.0	**71.4**
SimCC	256×192	✗	**25.7M**	**3.8**	70.8

Table 5. Ablation study of label smoothing. Results are achieved based on SimpleBaseline-Res50 [38] with the input size of 384×288 on COCO2017 val dataset. Employing label smoothing significantly improves the performance by 2.1 AP.

Label smoothing	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
w/o	71.3	88.8	78.2	67.8	78.2	77.3
equal	73.0	89.3	79.7	69.5	79.9	78.6
Gaussian	73.4	89.2	80.0	69.7	80.6	78.8
Laplace	73.0	89.3	79.7	69.3	80.3	78.4

counterpart w/ upsampling modules by **+0.4** AP, with **57.3%** fewer GFLOPs at the input size of 256×192.

Label smoothing. Label smoothing [32] is a commonly used strategy to improve generalization for the task of classification. To investigate its effect on our proposed method, we train SimpleBaseline-Res50 [38] based on SimCC with various label smoothing strategies: {*w/o*, *equal*, *Gaussian*, *Laplace*}. Table 5 demonstrates that label smoothing strategy does make a difference. Therefore, a promising way to further improve SimCC may be replacing the heuristic label smoothing strategy in a self-adaptive way. Further discussion is out the scope of this paper and we regard it as future work.

Table 6. Comparisons with 2D heatmap-based methods on CrowdPose test dataset. Results are achieved by HRNet-W32 [29] and "Heatmap" denotes 2D heatmap as an abbreviation. SimCC-based HRNet-W32 demonstrates consistent improvements compared to 2D heatmap-based methods.

Scheme	Input size	AP	AP^{50}	AP^{75}	AP^{E}	AP^{M}	AP^{H}
Heatmap	64×64	42.4	69.6	45.5	51.2	43.1	31.8
SimCC	64×64	**46.5**	**70.9**	**50.0**	**56.0**	**47.5**	**34.7**
Heatmap	256×192	66.4	81.1	71.5	74.0	67.4	55.6
SimCC	256×192	**66.7**	**82.1**	**72.0**	**74.1**	**67.8**	**56.2**

4.3 CrowdPose

One may concern about the performance of SimCC in dense pose scenes. Thus we further illustrate the effectiveness of the proposed SimCC on the CrowdPose [15] dataset, which contains much more crowded scenes than the COCO keypoint dataset. There are 20K images and 80K person instances in the CrowdPose. The training, validation and testing subset consist of about 10K, 2K, and 8K images respectively. Similar evaluation metric to that of COCO [19] is adopted here, with extra AP^E (AP scores on relatively easier samples) and AP^H (AP scores on harder samples). We follow the original paper [15] to adopt YoloV3 [26] as the human detector, and batch size is set as 64. We conduct comparison experiments on the CrowdPose test dataset, at the input size of 64×64 and 256×192 respectively. The results in Table 6 show that SimCC-based methods outperform heatmap-based counterparts.

4.4 MPII Human Pose Estimation

The MPII Human Pose dataset [1] contains 40K person samples with 16 joints labels. We point out that the data augmentation used on the MPII dataset is the same as that on COCO dataset.

Results on the validatoin set. We follow the evaluation procedure in HRNet [29]. The head-normalized probability of correct keypoint (PCKh) [1] score is used for model evaluation. The results are presented in Table 7. At the input

Table 7. Comparisons with 2D heatmap-based methods on the MPII validation set. Experiments are conducted based on HRNet-W32 [29] and "Heatmap" means 2D heatmap for short. "†" denotes that Gaussian label smoothing is utilized. Under stricter metric PCKh@0.1, SimCC shows clear gains across different input sizes.

Scheme	Input size	Hea	Sho	Elb	Wri	Hip	Kne	Ank	Mean
PCKh@0.5									
Heatmap	64×64	89.7	86.6	75.1	65.7	77.2	69.2	63.6	76.4
SimCC	64×64	**93.5**	**89.5**	**77.5**	**67.6**	**79.8**	**71.5**	**65.0**	**78.7**
Heatmap	256×256	97.1	95.9	90.3	**86.4**	89.1	**87.1**	**83.3**	**90.3**
SimCC	256×256	96.8	95.9	90.0	85.0	89.1	85.4	81.3	89.6
SimCC†	256×256	**97.2**	**96.0**	**90.4**	85.6	**89.5**	85.8	81.8	90.0
PCKh@0.1									
Heatmap	64×64	12.9	11.7	9.7	7.1	7.2	7.2	6.6	9.2
SimCC	64×64	**30.9**	**23.3**	**18.1**	**15.0**	**10.5**	**13.1**	**12.8**	**18.5**
Heatmap	256×256	44.5	37.3	37.5	36.9	15.1	25.9	27.2	33.1
SimCC	256×256	**50.1**	41.0	**45.3**	**42.4**	16.6	**29.7**	**30.3**	**37.8**
SimCC†	256×256	49.6	**41.9**	43.0	39.6	**17.0**	28.2	28.9	36.8

size of 256×256, SimCC-based methods achieve competitive performances under PCKh@0.5, and show clear gains under the stricter measurement PCKh@0.1.

5 Limitation and Future Work

SimCC introduced in this paper works under the setting of top-down human pose estimation. When it comes to bottom-up multi-person pose estimation, the presence of multiple people brings the identification ambiguity. Potential future work may introduce extra embeddings in a similar way to AE [22], in order to address the matching problem between candidate coordinate x and y values.

6 Conclusion

In this paper, we explore a simple yet promising coordinate representation (namely SimCC). It regards the keypoint localization task as two independent sub-tasks of classification for horizontal and vertical axes. The experimental results empirically show that the 2D structure might not be a key ingredient for coordinate representation to sustain superior performance. The proposed SimCC shows advantages over heatmap-based representation at model performances. Moreover, it may also inspire new works on lightweight model design for HPE.

Acknowledgements This work is supported in part by the National Natural Science Foundation of China under Grant 62171248, and the PCNL KEY project (PCL2021A07), and in part by the National Natural Science Foundation of China under Grant 61773117.

References

1. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on computer Vision and Pattern Recognition. pp. 3686–3693 (2014)
2. Cai, Y., Wang, Z., Luo, Z., Yin, B., Du, A., Wang, H., Zhang, X., Zhou, X., Zhou, E., Sun, J.: Learning delicate local representations for multi-person pose estimation. In: European Conference on Computer Vision. pp. 455–472. Springer (2020)
3. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: realtime multi-person 2d pose estimation using part affinity fields. IEEE transactions on pattern analysis and machine intelligence **43**(1), 172–186 (2019)
4. Carreira, J., Agrawal, P., Fragkiadaki, K., Malik, J.: Human pose estimation with iterative error feedback. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4733–4742 (2016)
5. Chen, T., Saxena, S., Li, L., Fleet, D.J., Hinton, G.: Pix2seq: A language modeling framework for object detection. arXiv preprint arXiv:2109.10852 (2021)
6. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7103–7112 (2018)
7. Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T.S., Zhang, L.: Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5386–5395 (2020)
8. Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: Rmpe: Regional multi-person pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2334–2343 (2017)
9. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
11. Huang, J., Zhu, Z., Guo, F., Huang, G.: 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. pp. 5700–5709 (2020)
12. Huang, S., Gong, M., Tao, D.: A coarse-fine network for keypoint localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3028–3037 (2017)
13. Kocabas, M., Karagoz, S., Akbas, E.: Multiposenet: Fast multi-person pose estimation using pose residual network. In: Proceedings of the European conference on computer vision (ECCV). pp. 417–433 (2018)
14. Li, J., Bian, S., Zeng, A., Wang, C., Pang, B., Liu, W., Lu, C.: Human pose regression with residual log-likelihood estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11025–11034 (2021)
15. Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H.S., Lu, C.: Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10863–10872 (2019)

16. Li, K., Wang, S., Zhang, X., Xu, Y., Xu, W., Tu, Z.: Pose recognition with cascade transformers. arXiv preprint arXiv:2104.06976 (2021)
17. Li, W., Wang, Z., Yin, B., Peng, Q., Du, Y., Xiao, T., Yu, G., Lu, H., Wei, Y., Sun, J.: Rethinking on multi-stage networks for human pose estimation. arXiv preprint arXiv:1901.00148 (2019)
18. Li, Y., Zhang, S., Wang, Z., Yang, S., Yang, W., Xia, S.T., Zhou, E.: Tokenpose: Learning keypoint tokens for human pose estimation. arXiv preprint arXiv:2104.03516 (2021)
19. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
20. Luo, Z., Wang, Z., Huang, Y., Tan, T., Zhou, E.: Rethinking the heatmap regression for bottom-up human pose estimation. arXiv preprint arXiv:2012.15175 (2020)
21. Mao, W., Ge, Y., Shen, C., Tian, Z., Wang, X., Wang, Z.: Tfpose: Direct human pose estimation with transformers. arXiv preprint arXiv:2103.15320 (2021)
22. Newell, A., Huang, Z., Deng, J.: Associative embedding: End-to-end learning for joint detection and grouping. In: Advances in Neural Information Processing Systems (2017)
23. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European conference on computer vision. pp. 483–499. Springer (2016)
24. Nie, X., Feng, J., Zhang, J., Yan, S.: Single-stage multi-person pose machines. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6951–6960 (2019)
25. Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., Murphy, K.: Towards accurate multi-person pose estimation in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4903–4911 (2017)
26. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
27. Rezende, D., Mohamed, S.: Variational inference with normalizing flows. In: International conference on machine learning. pp. 1530–1538. PMLR (2015)
28. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
29. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5693–5703 (2019)
30. Sun, X., Shang, J., Liang, S., Wei, Y.: Compositional human pose regression. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2602–2611 (2017)
31. Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral human pose regression. In: Proceedings of the European Conference on Computer Vision. pp. 529–545 (2018)
32. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
33. Tian, Z., Chen, H., Shen, C.: Directpose: Direct end-to-end multi-person pose estimation. arXiv preprint arXiv:1911.07451 (2019)
34. Tompson, J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. arXiv preprint arXiv:1406.2984 (2014)

35. Toshev, A., Szegedy, C.: Deeppose: Human pose estimation via deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1653–1660 (2014)
36. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017)
37. Wei, F., Sun, X., Li, H., Wang, J., Lin, S.: Point-set anchors for object detection, instance segmentation and pose estimation. In: European Conference on Computer Vision. pp. 527–544. Springer (2020)
38. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: Proceedings of the European conference on computer vision. pp. 466–481 (2018)
39. Xiong, Y., Zhou, Z., Dou, Y., Su, Z.: Gaussian vector: An efficient solution for facial landmark detection. In: Proceedings of the Asian Conference on Computer Vision (2020)
40. Yang, S., Quan, Z., Nie, M., Yang, W.: Transpose: Towards explainable human pose estimation by transformer. arXiv preprint arXiv:2012.14214 (2020)
41. Yang, W., Li, S., Ouyang, W., Li, H., Wang, X.: Learning feature pyramids for human pose estimation. In: proceedings of the IEEE international conference on computer vision. pp. 1281–1290 (2017)
42. Yin, S., Wang, S., Chen, X., Chen, E.: Attentive one-dimensional heatmap regression for facial landmark detection and tracking (2020)
43. Zhang, F., Zhu, X., Dai, H., Ye, M., Zhu, C.: Distribution-aware coordinate representation for human pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7093–7102 (2020)
44. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv preprint arXiv:1904.07850 (2019)