

# Joint COCO and Mapillary Workshop at ICCV 2019: COCO 2019 Keypoint Detection Challenge Track

## Technical Report: Distribution-Aware Coordinate Representation of Keypoint for Human Pose Estimation

Hanbin Dai<sup>1\*</sup>      Liangbo Zhou<sup>1\*</sup>      Feng Zhang<sup>1\*</sup>  
Zhengyu Zhang<sup>2\*</sup>      Hong Hu<sup>1\*</sup>      Xiatian Zhu<sup>3\*</sup>      Mao Ye<sup>1</sup>  
University of Electronic Science and Technology of China<sup>1</sup>  
Shenzhen University<sup>2</sup>      University of Surrey<sup>3</sup>  
{daihanbin.ac, heathhose96, zhangfengwcy, Andrewhuhong}@gmail.com  
zhengyuzhang23@outlook.com      eddy.zhuxt@gmail.com      cvlab.uestc@gmail.com

### Abstract

*In this paper, we focus on the coordinate representation in human pose estimation. While being the standard choice, heatmap based representation has not been systematically investigated. We found that the process of coordinate decoding (i.e. transforming the predicted heatmaps to the coordinates) is surprisingly significant for human pose estimation performance, which nevertheless was not recognised before. In light of the discovered importance, we further probe the design limitations of the standard coordinate decoding method and propose a principled distribution-aware decoding method. Meanwhile, we improve the standard coordinate encoding process (i.e. transforming ground-truth coordinates to heatmaps) by generating accurate heatmap distributions for unbiased model training. Taking them together, we formulate a novel Distribution-Aware coordinate Representation for Keypoint (DARK) method. Serving as a model-agnostic plug-in, DARK significantly improves the performance of a variety of state-of-the-art human pose estimation models. Extensive experiments show that DARK yields the best results on COCO keypoint detection challenge, validating the usefulness and effectiveness of our novel coordinate representation idea. The project page containing more details is at <https://ilovepose.github.io/coco/>*

### 1. Introduction

Human pose estimation is a challenging problem in computer vision aiming at finding the coordinates of hu-

man body parts. Recently, convolutional neural networks (CNNs) have achieved significant success [10, 6, 11, 5, 13, 12, 8]. However, these methods typically focus on designing pose specific architecture, ignoring the coordinate representation of body parts. In the classification task, the one-hot vectors are utilised to represent the object class, so that the model can learn the target easily. A human pose estimation model also needs a target representation (coordinate encoding and decoding). The *de facto* standard coordinate representation of body part is coordinate heatmap generated using a 2D Gaussian distribution/kernel centred at the labelled coordinate of each joint [9]. Down-sampling is often needed for controlling the computational cost.

In the literature, the problem of coordinate encoding and decoding (i.e. denoted as coordinate representation) gains little attention, although being indispensable in model training and inference. Contrary to the existing human pose estimation studies, in this work we dedicatedly investigate the problem of joint coordinate representation including encoding and decoding. Moreover, we recognise that the heatmap resolution is one major obstacle that prevents the use of smaller input resolution for faster model inference. In light of the discovered significance of coordinate representation, we conduct in-depth investigation and recognise that one key limitation lies in the coordinate decoding process. Whilst existing standard shifting operation has shown to be effective as found in this study, we propose a principled distribution-aware representation method for more accurate joint localisation at sub-pixel accuracy. Specifically, it is designed to comprehensively account for the distribution information of heatmap activation via Taylor-expansion based distribution approximation. Besides, we observe that the standard method for generating the ground-truth heatmaps

---

\*equal contribution

suffers from quantisation/discretisation errors, leading to imprecise supervision and inferior performance. To solve this issue, we propose generating unbiased heatmaps allowing Gaussian kernel being centred at sub-pixel locations.

The contribution of this work is that, we discover the previously unrealised significance of coordinate representation in human pose estimation, and propose a novel Distribution-Aware coordinate Representation for Keypoint (DARK) method with two key components: (1) efficient Taylor-expansion based coordinate decoding, and (2) unbiased sub-pixel centred coordinate encoding. Importantly, existing human pose methods can be seamlessly benefited from DARK without any algorithmic modification. Extensive experiments on COCO keypoint benchmark show that our method provides significant performance gain for the existing state-of-the-art human pose estimation model [5, 12, 8], achieving the best single model accuracy. DARK favourably enables the use of smaller input image resolutions with much smaller performance degradation, whilst dramatically boosting the model inference efficiency.

## 2. Method

### 2.1. Human Pose Estimator

We find a significant performance bottleneck in the coordinate representation (coordinate encoding and decoding), and introduce a principled solution, named as *Distribution-Aware coordinate Representation for Keypoint* (DARK). In the following we first describe the decoding process, focusing on the limitation analysis of the existing standard method and the development of a novel solution. Then, we discuss and address the limitations of the encoding process.

#### 2.1.1 Coordinate Decoding

Suppose a pose estimator outputs a heatmap matching the spatial size of an input image. It is easy to obtain the location of the body joints by identifying the maximum activation in the heatmap. However, this is often not the case due to the computation budget constraint. Instead, we need to upscale the low-resolution heatmap to the original image resolution. This involves a sub-pixel localisation problem. The standard method is to offset the max-activation prediction by a quarter of a pixel in the direction towards the second max activation before transforming back to the original coordinate space of the input image. This hand-designed method is not sufficiently accurate without good insights.

To solve the sub-pixel localisation problem, we propose a Taylor-expansion based re-localisation method, called distribution-aware maximum re-localisation (Fig 1 (b)). Specifically, we exploit the Taylor-expansion theory to estimate the underlying max activation in a Gaussian distribution assumption. The predicted Gaussian heatmap is often

ill-conditioned which may hurt the offset estimation. We therefore further design a heatmap distribution modulation method (Fig 1(a)) for preprocessing. Specifically, a Gaussian kernel is utilised to smooth the predicted heatmap.

#### 2.1.2 Coordinate Encoding

The heatmap based representation assume the coordinate of a body part follows a 2D Gaussian distribution. In the coordinate encoding phase, the original person images is down-sampled into the model input size. So, the ground-truth joint coordinates require to be transformed accordingly before generating the heatmaps.

Formally, we denote by  $\mathbf{g} = (u, v)$  the ground-truth coordinate of a joint. The resolution reduction is defined as:

$$\mathbf{g}' = (u', v') = \frac{\mathbf{g}}{\lambda} = \left(\frac{u}{\lambda}, \frac{v}{\lambda}\right) \quad (1)$$

where  $\lambda$  is the downsampling ratio.

Conventionally, for facilitating the kernel generation, we often quantise  $\mathbf{g}'$ :

$$\mathbf{g}'' = (u'', v'') = \text{quantise}(\mathbf{g}') = \text{quantise}\left(\frac{u}{\lambda}, \frac{v}{\lambda}\right) \quad (2)$$

where  $\text{quantise}()$  specifies a quantisation function, with the common choices including floor, ceil and round.

Subsequently, the heatmap centred at the quantised coordinate  $\mathbf{g}''$  can be synthesised through:

$$\mathcal{G}(x, y; \mathbf{g}'') = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(x - u'')^2 + (y - v'')^2}{2\sigma^2}\right) \quad (3)$$

where  $(x, y)$  specifies a pixel location in the heatmap, and  $\sigma$  denotes a fixed spatial variance.

Obviously, the heatmaps generated in the above way are *inaccurate* and *biased* due to the quantisation error. This may introduce sub-optimal supervision signals and result in degraded model performance, particularly for the case of accurate coordinate encoding as proposed in this work.

To address this issue, we simply place the heatmap centre at the non-quantised location  $\mathbf{g}'$  which represents the *accurate* ground-truth coordinate. We still apply Eq. (3) but replacing  $\mathbf{g}''$  with  $\mathbf{g}'$ .

### 2.2. Person Detection

To obtain good person detection results efficiently, we use the Hybrid Task Cascade (HTC) detector [2] and the SNIPER detector [7] jointly for the challenge entry model.

## 3. Experiments

### 3.1. Datasets

We used two datasets. (1) The COCO keypoint dataset [4] consists of about 200K images containing 250K person

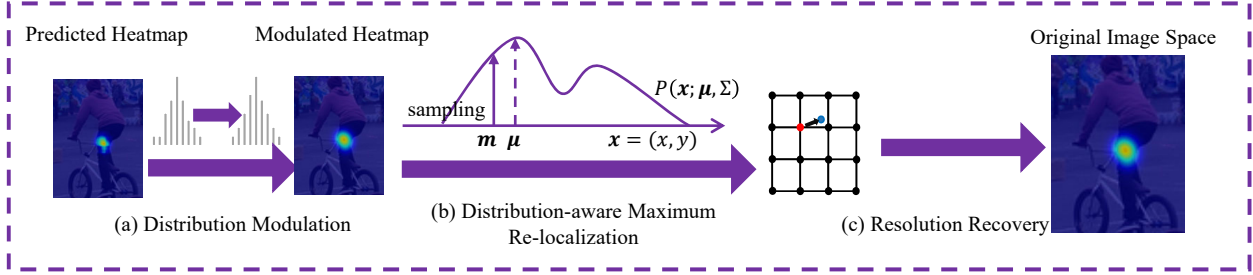


Figure 1: Overview of the proposed distribution aware coordinate decoding method.

Table 1: Effect of coordinate decoding on COCO val. Model: HRNet-W32; Input size:  $128 \times 96$ .

Decoding	$AP$	$AP^{50}$	$AP^{75}$	$AP^M$	$AP^L$	$AR$
No Shifting	61.2	88.1	72.3	59.0	66.3	68.7
Standard Shifting	66.9	<b>88.7</b>	76.3	64.6	72.3	73.7
<b>Ours</b>	<b>68.4</b>	88.6	<b>77.4</b>	<b>66.0</b>	<b>74.0</b>	<b>74.9</b>

Table 2: Effect of coordinate encoding on COCO val. Model: HRNet-W32; Input size:  $128 \times 96$ .

Encode	Decode	$AP$	$AP^{50}$	$AP^{75}$	$AP^M$	$AP^L$	$AR$
Biased	Standard	66.9	88.7	76.3	64.6	72.3	73.7
<b>Unbiased</b>	Standard	<b>68.0</b>	<b>88.9</b>	<b>77.0</b>	<b>65.4</b>	<b>73.7</b>	<b>74.5</b>
Biased	<b>Ours</b>	68.4	88.6	77.4	66.0	74.0	74.9
<b>Unbiased</b>	<b>Ours</b>	<b>70.7</b>	<b>88.9</b>	<b>78.4</b>	<b>67.9</b>	<b>76.6</b>	<b>76.7</b>

instances labelled with 17 joints. It has four splits: train, val, test-dev, test-challenge with 118K, 5K, 20K and 20K images respectively. (2) The AIC dataset [1] contains about 300k images and 700k person instance labelled with 14 key-points. It has four splits: train, val, test A, test B with 210K, 30K, 30K and 30K images respectively.

## 3.2. Ablation Study

### 3.2.1 Evaluating Coordinate Representation

In this test, we used the person detection results from [8]. By default we used HRNet-W32 as the backbone model and  $128 \times 96$  as the input size, and reported the accuracy results on the COCO validation set.

(i) **Coordinate decoding** We evaluated the proposed coordinate decoding. The conventional biased heatmaps were used. We compared the proposed distribution-aware shifting method with *no shifting* (*i.e.* directly using the maximal activation location), and the *standard shifting* in [5, 3, 12, 8]. We observed in Table 1 that: (i) The standard shifting gives as high as 5.7% AP accuracy boost, which is surprisingly effective. This reveals previously unseen significance of coordinate decoding to human pose estimation. (ii) Despite the great gain by the standard decoding method, the proposed model further improves AP score by 1.5%.

(ii) **Coordinate encoding** We compared the proposed *unbiased* encoding with the standard *biased* encoding, along with both the standard and our decoding method. We observed from Table 2 that our unbiased encoding with accurate kernel centre brings positive performance margin, regardless of the coordinate decoding method.

Table 3: Effect of input image size on COCO val. DARK uses HRNet-W32 (HRN32) as backbone.

Method	Input size	GFLOPs	$AP$	$AP^{50}$	$AP^{75}$	$AP^M$	$AP^L$	$AR$
HRN32	$128 \times 96$	1.8	66.9	88.7	76.3	64.6	72.3	73.7
<b>DARK</b>			<b>70.7</b>	<b>88.9</b>	<b>78.4</b>	<b>67.9</b>	<b>76.6</b>	<b>76.7</b>
HRN32	$256 \times 192$	7.1	74.4	<b>90.5</b>	81.9	70.8	81.0	79.8
<b>DARK</b>			<b>75.6</b>	<b>90.5</b>	<b>82.1</b>	<b>71.8</b>	<b>82.8</b>	<b>80.8</b>
HRN32	$384 \times 288$	16.0	75.8	90.6	82.5	72.0	82.7	80.9
<b>DARK</b>			<b>76.6</b>	<b>90.7</b>	<b>82.8</b>	<b>72.7</b>	<b>83.9</b>	<b>81.5</b>

(iii) **Input resolution** We examined the impact of input image resolution/size. We compared our DARK model (HRNet-W32 as backbone) with the original HRNet-W32 using the biased heatmap supervision for training and the standard shifting for testing. From Table 3 we have a couple of observations: (a) With reduced input image size, as expected the model performance consistently degrades whilst the inference cost drops clearly. (b) With the support of DARK, the model performance loss can be effectively mitigated, especially in case of very small input resolution (*i.e.* very fast model inference).

### 3.2.2 Effect of DARK

We further evaluate the effect of DARK on the COCO test-dev set. We compared the HRNet-W48(HRN48) with DARK using HRNet-W48 backbone. We observed from Table 4 that DARK gives a clear performance gain.

Table 4: Effect of DARK on COCO test-dev; Input size:  $384 \times 288$ ; Training data: COCO train; Detection: MSRA[8].

Method	$AP$	$AP^{50}$	$AP^{75}$	$AP^M$	$AP^L$	$AR$
HRN48	75.5	<b>92.5</b>	83.3	71.9	81.5	80.5
DARK	<b>76.2</b>	<b>92.5</b>	<b>83.6</b>	<b>72.5</b>	<b>82.4</b>	<b>81.1</b>

Table 5: Effect of extra training data on COCO test-dev; Model: DARK using HRNet-W48 as backbone; Input size:  $384 \times 288$ ; Detection: MSRA[8].

Dataset	$AP$	$AP^{50}$	$AP^{75}$	$AP^M$	$AP^L$	$AR$
COCO	76.2	92.5	83.6	72.5	82.4	81.1
COCO+AIC	<b>77.4</b>	<b>92.6</b>	<b>84.6</b>	<b>73.6</b>	<b>83.7</b>	<b>82.3</b>

Table 6: Effect of person detection on COCO test-dev; Model: DARK using HRNet-W48 as backbone; Input size:  $384 \times 288$ ; Training data: COCO train+AIC train; Detection: MSRA[8], HTC[2], SNIPER[7].

Detector	$AP$	$AP^{50}$	$AP^{75}$	$AP^M$	$AP^L$	$AR$
MSRA	77.4	92.6	84.6	73.6	83.7	82.3
SNIPER	78.0	<b>93.6</b>	85.1	74.2	83.6	82.6
HTC	<b>78.2</b>	93.5	<b>85.5</b>	<b>74.4</b>	83.7	83.2
HTC+SNIPER	<b>78.2</b>	93.5	<b>85.5</b>	<b>74.4</b>	<b>84.2</b>	<b>83.5</b>

### 3.2.3 Effect of Extra Training Data

We examined the impact of extra training data with DARK(HRNet-W48) from AIC. The results in Table 5 show that extra training data brings a positive performance boost, as expected.

### 3.2.4 Effect of Person Detection

We examined different person detectors. We observed in Table 6 that: (i) HTC is the best detector; (ii) the combined detection can boost the overall performance.

### 3.2.5 Effect of Model Ensemble

We formed two ensembles: one with 3 models and one with 8 models. They were trained by DARK with varying backbones (HRNet-W48, HRNet-W32, ResNet-152), training data (COCO, COCO+AIC), and batch sizes (small and large). Table 7 shows that model ensemble helps.

## 3.3. ICCV Keypoint Detection Challenge

We used an ensemble of 8 DARK models for the challenge. Table 8 shows that our method achieves 76.4% AP

Table 7: Best single model vs. ensemble of 3/8 models on COCO test-dev; Input size:  $384 \times 288$ ; Detection: HTC+SNIPER;

Model	$AP$	$AP^{50}$	$AP^{75}$	$AP^M$	$AP^L$	$AR$
Best Single	78.2	93.5	85.5	74.4	84.2	83.5
Ensemble(3)	78.7	93.6	<b>86.0</b>	74.7	84.3	<b>83.5</b>
Ensemble(8)	<b>78.9</b>	<b>93.8</b>	<b>86.0</b>	<b>75.1</b>	<b>84.4</b>	<b>83.5</b>

Table 8: Result of 8-model ensemble on COCO test-challenge.

$AP$	$AP^{50}$	$AP^{75}$	$AP^M$	$AP^L$	$AR$
76.4	92.5	82.7	70.9	83.8	81.6

for multi-person pose estimation on COCO test-challenge set.

## 4. Conclusion

We presented a strong human pose estimation method based on a novel distribution-aware coordinate representation idea. It achieves very competitive results on COCO keypoint detection challenge. Please visit our project page for more details.

## References

- [1] Ai challenger human pose estimation dataset. <https://challenger.ai/competition/keypoint>. 3
- [2] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Hybrid task cascade for instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2, 4
- [3] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2018. 3
- [4] 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*, 2014. 2
- [5] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, 2016. 1, 2, 3
- [6] Tomas Pfister, James Charles, and Andrew Zisserman. Flowing convnets for human pose estimation in videos. In *IEEE International Conference on Computer Vision*, 2015. 1
- [7] Bharat Singh, Mahyar Najibi, and Larry S Davis. SNIPER: Efficient multi-scale training. *Advances in Neural Information Processing Systems*, 2018. 2, 4
- [8] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose esti-

- mation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. [1](#), [2](#), [3](#), [4](#)
- [9] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in Neural Information Processing Systems*, 2014. [1](#)
- [10] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. [1](#)
- [11] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. [1](#)
- [12] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *European Conference on Computer Vision*, 2018. [1](#), [2](#), [3](#)
- [13] Wei Yang, Shuang Li, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Learning feature pyramids for human pose estimation. In *IEEE International Conference on Computer Vision*, 2017. [1](#)