Distribution-Aware Coordinate Representation of Keypoint for Human Pose Estimation

Feng Zhang[1] **Xiatian Zhu**[2] **Mao Ye**[1] **Ce Zhu**[1]
University of Electronic Science and Technology of China[1] University of Surrey[2]
{zhangfengwcy, eddy.zhuxt, cvlab.uestc}@gmail.com eczhu@uestc.edu.cn

Abstract

While being the *de facto* standard coordinate representation in human pose estimation, *heatmap* is never systematically investigated in the literature, to our best knowledge. This work fills this gap by studying the coordinate representation with a particular focus on the heatmap. Interestingly, we found that the process of *decoding* the predicted heatmaps into the final joint coordinates in the original image space 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 widely used by existing methods, and propose a more 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 the two together, we formulate a novel *Distribution-Aware coordinate Representation of 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 two common benchmarks, MPII and COCO, consistently validating the usefulness and effectiveness of our novel coordinate representation idea. The project page is at https://ilovepose.github.io/coco/

Introduction

Human pose estimation is a fundamental computer vision problem that aims to detect the *spatial location* (*i.e. coordinate*) of human body joints in unconstrained images (Andriluka et al. 2014). It is a non-trivial task as the appearance of body joints vary dramatically due to diverse styles of clothes, arbitrary occlusion, and unconstrained background contexts, whilst it is needed to identify the *fine-grained* joint coordinates. As strong image processing models, convolutional neural networks (CNNs) excel at this task (LeCun et al. 1998). Existing works typically focus on designing the CNN architecture tailored particularly for human pose inference (Newell, Yang, and Deng 2016; Sun et al. 2019).

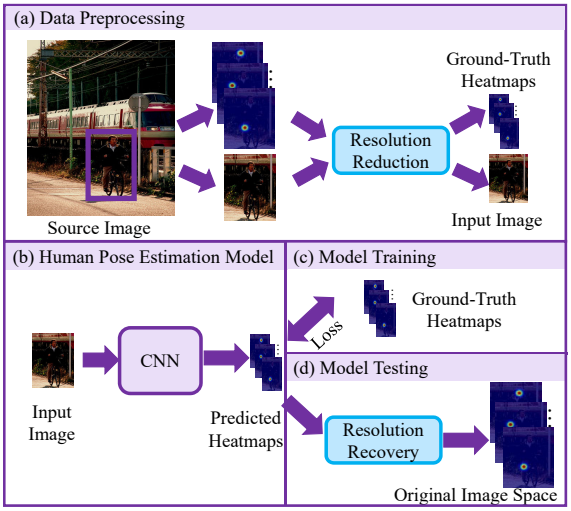

Figure 1: Pipeline of a human pose estimation system. For efficiency, *resolution reduction* is often applied on the original person detection bounding boxes as well as the ground-truth heatmap supervision. That is, the model operates in a low-resolution image space. At test time, a corresponding *resolution recovery* is therefore necessary in order to obtain the joint coordinate prediction in the original image space.

Analogous to the common *one-hot vectors* as the object class label representation in image classification, a human pose CNN model also requires a ***label representation*** for encoding the *body joint coordinate labels*, so that the supervised learning loss can be quantified and computed during training and the joint coordinates can be inferred properly[1]. The *de facto* standard label representation is *coordinate heatmap*, generated as a 2-dimensional Gaussian distribution/kernel centred at the labelled coordinate of each joint (Tompson et al. 2014). It is obtained from a ***coordinate encoding*** process, *from coordinate to heatmap*. Heatmap is

[1]The *label representation* is for encoding the label annotations (*e.g.* 1,000 one-hot vectors for 1,000 object class labels in ImageNet), totally different from the *data representation* for encoding the data samples (*e.g.* the object images from ImageNet).

characterised by giving spatial support around the ground-truth location, considering not only the contextual clues but also the inherent target position ambiguity. Importantly, this may effectively reduce the model overfitting risk in training, in a similar spirit of the class label smoothing regularisation (Szegedy et al. 2016). Come as no surprise, the state-of-the-art pose models (Newell, Yang, and Deng 2016; Xiao, Wu, and Wei 2018; Sun et al. 2019) are based on the heatmap coordinate representation.

With the heatmap label representation, one major obstacle is that, the computational cost is a *quadratic* function of *the input image resolution*, preventing the CNN models from processing the typically *high-resolution* raw imagery data. To be computationally affordable, a standard strategy (see Fig. 1) is to downsample all the person bounding box images at arbitrarily large resolutions into a prefixed small resolution with a data preprocessing procedure, before being fed into a human pose estimation model. Aiming to predict the joint location in the *original* image coordinate space, after the heatmap prediction a corresponding *resolution recovery* is required for transforming back to the original coordinate space. The final prediction is considered as the location with the maximal activation. We call this process as ***coordinate decoding***, *from heatmap to coordinate*. It is worthy noting that quantisation error can be introduced during the above resolution reduction. To alleviate this problem, during the existing coordinate decoding process a hand-crafted shifting operation is usually performed according to the direction from the highest activation to the second highest activation (Newell, Yang, and Deng 2016).

In the literature, the problem of coordinate encoding and decoding (*i.e.* denoted as *coordinate representation*) gains little attention, although being indispensable in model inference. In contrast to the current research focus on designing more effective CNN structures, we reveal a *surprisingly* important role the coordinate representation plays on the model performance, much more significant than expected. For instance, with the state-of-the-art model HRNet-W32 (Sun et al. 2019), the aforementioned shifting operation of coordinate encoding brings as high as 5.7% AP on the challenging COCO validation set (Table 1). It is noteworthy to mention that, this gain is already much more significant than those by most individual art methods. But it is never well noticed and carefully investigated in the literature to our best knowledge.

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. When decreasing the input resolution from 256×192 to 128×96, the model performance of HRNet-W32 drops significantly from 74.4% to 66.9% on the COCO validation set, although the model inference cost falls from 7.1×10^9 to 1.8×10^9 FLOPs.

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 suffers from *quantisation errors*, leading to imprecise supervision signals and inferior model performance. To solve this issue, we propose generating the *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 of 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 two common benchmarks (MPII and COCO) show that our method provides significant performance improvement for existing state-of-the-art human pose estimation models (Sun et al. 2019; Xiao, Wu, and Wei 2018; Newell, Yang, and Deng 2016), achieving the best single model accuracy on COCO and MPII. DARK favourably enables the use of smaller input image resolutions with much smaller performance degradation, whilst dramatically boosting the model inference efficiency therefore facilitating low-latency and low-energy applications as required in embedded AI scenarios.

Related Work

There are two common coordinate representation designs in human pose estimation: direct coordinate and heatmap. Both are used as the regression targets for model training.

Coordinate regression Directly taking the coordinates as model output target is straightforward and intuitive. But only a handful of existing methods adopt this design (Toshev and Szegedy 2014; Fan et al. 2015; Carreira et al. 2016; Sun et al. 2018). One plausible reason is that, this representation lacks the spatial and contextual information, making the learning of human pose model extremely challenging due to the intrinsic visual ambiguity in joint location.

Heatmap regression The heatmap representation elegantly addresses the above limitations. It was firstly introduced in (Tompson et al. 2014) and rapidly became the most commonly used coordinate representation. Generally, the mainstream research focus is on designing network architectures for more effectively regressing the heatmap supervision. Representative design improvements include sequential modelling (Gkioxari, Toshev, and Jaitly 2016; Belagiannis and Zisserman 2017), receptive field expansion (Wei et al. 2016), position voting (Lifshitz, Fetaya, and Ullman 2016), intermediate supervision (Newell, Yang, and Deng 2016; Wei et al. 2016), pairwise relations modelling (Chen and Yuille 2014), tree structure modelling (Chu et al. 2016b; Yang et al. 2016; Chu et al. 2016a; Sun et al. 2017; Tang, Yu, and Wu 2018), pyramid residual learning (Yang

et al. 2017), cascaded pyramid learning (Chen et al. 2018), knowledge-guided learning (Ning, Zhang, and He 2017), active learning (Liu and Ferrari 2017), adversarial learning (Chen et al. 2017), deconvolution upsampling (Xiao, Wu, and Wei 2018), multi-scale supervision (Ke et al. 2018), attentional mechanism (Liu et al. 2018; Su et al. 2019), and high-resolution representation preserving (Sun et al. 2019).

In contrast to all previous works, we instead investigate the issues of heatmap representation on human pose estimation, a largely ignored perspective in the literature. Not only do we reveal a big impact of resolution reduction in the process of using heatmap but also we propose a principled coordinate representation method for significantly improving the performance of existing models. Crucially, our method can be seamlessly integrated without model design change.

Methodology

We consider the coordinate representation problem including encoding and decoding in human pose estimation. The objective is to predict the joint coordinates in a given input image. To that end, we need to learn a regression model from the input image to the output coordinates, and the *heatmap* is often leveraged as coordinate representation during both model training and testing. Specifically, we assume access to a training set of images. To facilitate the model learning, we *encode* the labelled ground-truth coordinate of a joint into a heatmap as the supervised learning target. During testing, we then need to *decode* the predicted heatmap into the coordinate in the original image coordinate space.

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 further discuss and address the limitations of the encoding process. Lastly, we describe the integration of existing human pose estimation models with the proposed method.

Coordinate Decoding

Despite being considered as an insignificant component of the model testing pipeline, as we found in this study, coordinate decoding turns out to be one of the most significant performance contributors for human pose estimation in images (see Table 1). Specifically, this is a process of translating a predicted heatmap of each individual joint into a coordinate in the *original* image space. Suppose the heatmap has the same spatial size as the original image, we only need to find the location of the maximal activation as the joint coordinate prediction, which is straightforward and simple. However, this is often not the case as interpreted above. Instead, we need to upsample the heatmaps to the original image resolution by a sample-specific unconstrained factor $\lambda \in \mathcal{R}_+$. This involves a *sub-pixel localisation* problem. Before introducing our method, we first revisit the standard coordinate decoding method used in existing pose estimation models.

The standard coordinate decoding method is designed empirically according to model performance (Newell, Yang, and Deng 2016). Specifically, given a heatmap $\boldsymbol{h}$ predicted by a trained model, we first identify the coordinates of the maximal ($\boldsymbol{m}$) and second maximal ($\boldsymbol{s}$) activation. The joint location is then predicted as

$$\boldsymbol{p} = \boldsymbol{m} + 0.25\frac{\boldsymbol{s}-\boldsymbol{m}}{\|\boldsymbol{s}-\boldsymbol{m}\|_2} \tag{1}$$

where $\|\cdot\|_2$ defines the magnitude of a vector. This means that the prediction is as the maximal activation with a 0.25 pixel (*i.e.* sub-pixel) shifting towards the second maximal activation in the heatmap space. The final coordinate prediction in the original image is computed as:

$$\hat{\boldsymbol{p}} = \lambda \boldsymbol{p} \tag{2}$$

where λ is the resolution reduction ratio.

Remarks The aim of the sub-pixel shifting in Eq. (1) is to compensate the quantisation effect of image resolution downsampling. That being said, the maximum activation in the predicted heatmap does not correspond to the accurate position of the joint in the original coordinate space, but only to a *coarse* location. As we will show, this shifting *surprisingly* brings a significant performance boost (Table 1). This may partly explain why it is often used as a standard operation in model test. Interestingly, to our best knowledge no specific work has delved into the effect of this operation on human pose estimation performance. Therefore, its true significance has never been really recognised and reported in the literature. While this standard method lacks intuition and interpretation in design, no dedicated investigation has been carried out for improvement. We fill this gap by presenting a principled method for shifting estimation and finally more accurate human pose estimation.

The proposed coordinate decoding method explores the distribution structure of the predicted heatmap to infer the underlying maximum activation. This differs dramatically to the standard method above relying on a hand-designed offset prediction, with little design justification and rationale.

Specifically, to obtain the accurate location at the degree of sub-pixel, we assume the predicted heatmap follows a 2D Gaussian distribution, same as the ground-truth heatmap. Therefore, we represent the predicted heatmap as

$$\mathcal{G}(\boldsymbol{x};\boldsymbol{\mu},\Sigma) = \frac{1}{(2\pi)|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) \tag{3}$$

where $\boldsymbol{x}$ is a pixel location in the predicted heatmap, $\boldsymbol{\mu}$ is the Gaussian mean (centre) corresponding to the *to-be-estimated* joint location. The covariance Σ is a diagonal matrix, same as that used in coordinate encoding:

$$\Sigma = \begin{bmatrix}\sigma^2 & 0\\ 0 & \sigma^2\end{bmatrix} \tag{4}$$

where σ is the standard deviation same for both directions.

In the log-likelihood optimisation principle (Goodfellow, Bengio, and Courville 2016), we transform $\mathcal{G}$ through logarithm to facilitate inference while keeping the original location of the maximum activation as:

$$\begin{aligned}\mathcal{P}(\boldsymbol{x};\boldsymbol{\mu},\Sigma) = \ln(\mathcal{G}) = &-\ln(2\pi) - \frac{1}{2}\ln(|\Sigma|)\\ &-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\end{aligned} \tag{5}$$

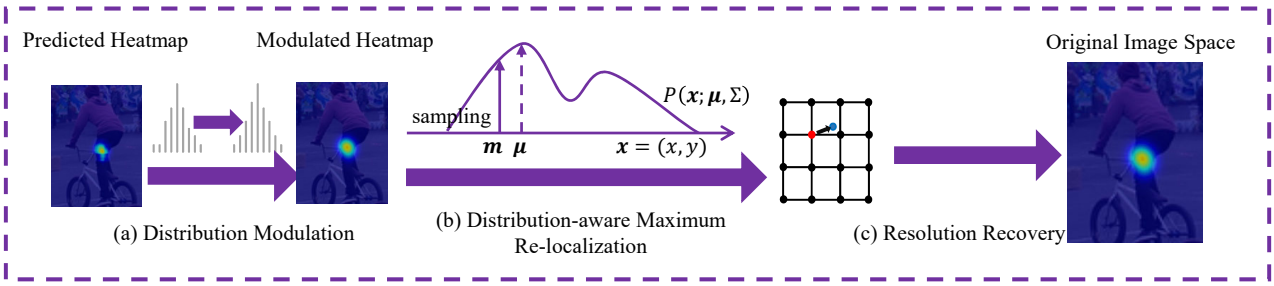

Figure 2: Overview of the proposed distribution aware coordinate decoding method.

Our objective is to estimate $\boldsymbol{\mu}$. As an extreme point in the distribution, it is well-known that the first derivative at the location $\boldsymbol{\mu}$ meets a condition as:

$$\mathcal{D}'(\boldsymbol{x})\Big|_{\boldsymbol{x}=\boldsymbol{\mu}} = \frac{\partial \mathcal{P}^T}{\partial \boldsymbol{x}}\Big|_{\boldsymbol{x}=\boldsymbol{\mu}} = -\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\Big|_{\boldsymbol{x}=\boldsymbol{\mu}} = 0 \quad (6)$$

To explore this condition, we adopt the Taylor's theorem. Formally, we approximate the activation $\mathcal{P}(\mu)$ by a Taylor series (up to the quadratic term) evaluated at the maximal activation $\boldsymbol{m}$ of the predicted heatmap as

$$\mathcal{P}(\boldsymbol{\mu}) = \mathcal{P}(\boldsymbol{m}) + \mathcal{D}'(\boldsymbol{m})(\boldsymbol{\mu}-\boldsymbol{m}) + \frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{m})^T\mathcal{D}''(\boldsymbol{m})(\boldsymbol{\mu}-\boldsymbol{m}) \quad (7)$$

where $\mathcal{D}''(\boldsymbol{m})$ denotes the second derivative (*i.e.* Hessian) of $\mathcal{P}$ evaluated at $\boldsymbol{m}$, formally defined as:

$$\mathcal{D}''(\boldsymbol{m}) = \mathcal{D}''(\boldsymbol{x})\Big|_{\boldsymbol{x}=\boldsymbol{m}} = -\Sigma^{-1} \quad (8)$$

The intuition of selecting $\boldsymbol{m}$ to approximate $\boldsymbol{\mu}$ is that it represents a good coarse joint prediction that approaches $\boldsymbol{\mu}$.

Taking Eq. (6), (7), and (8) together, we eventually obtain

$$\boldsymbol{\mu} = \boldsymbol{m} - \big(\mathcal{D}''(\boldsymbol{m})\big)^{-1}\mathcal{D}'(\boldsymbol{m}) \quad (9)$$

where $\mathcal{D}''(\boldsymbol{m})$ and $\mathcal{D}'(\boldsymbol{m})$ can be estimated efficiently from the heatmap. Once obtaining $\boldsymbol{\mu}$, we also apply Eq. (2) to predict the coordinate in the original image space.

Remarks In contrast to the standard method considering the second maximum activation alone in heatmap, the proposed coordinate decoding fully explores the heatmap distributional statistics for revealing the underlying maximum more accurately. In theory, our method is based on a principled distribution approximation under a training-supervision-consistent assumption that the heatmap is in a Gaussian distribution. Crucially, it is very efficient computationally as it only needs to compute the first and second derivative of one location per heatmap. Consequently, existing human pose estimation approaches can be readily benefited without any computational cost barriers.

Heatmap distribution modulation As the proposed coordinate decoding method is based on a Gaussian distribution assumption, it is necessary for us to examine how well this condition is satisfied. We found that, often, the heatmaps predicted by a human pose estimation model do *not* exhibit good-shaped Gaussian structure compared to the training heatmap data. As shown in Fig. 3(a), the heatmap usually presents multiple peaks around the maximum activation. This may cause negative effects to the performance of our decoding method. To address this issue, we propose *modulating* the heatmap distribution beforehand.

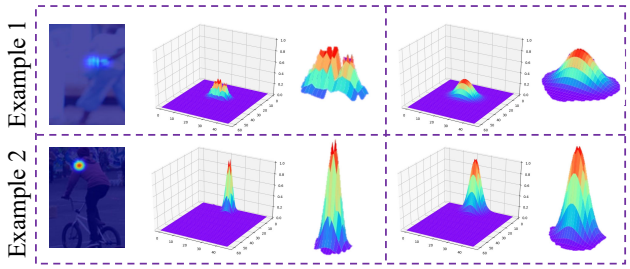

Figure 3: Illustration of heatmap distribution modulation. (a) Predicted heatmap; (b) Modulated heatmap distribution.

Specifically, to match the requirement of our method we propose exploiting a Gaussian kernel K with the same variation as the training data to smooth out the effects of multiple peaks in the heatmap $\boldsymbol{h}$, formally as

$$\boldsymbol{h}' = K \circledast \boldsymbol{h} \quad (10)$$

where $\circledast$ specifies the convolution operation.

To preserve the original heatmap's magnitude, we finally scale $\boldsymbol{h}'$ so that its maximum activation is equal to that of $\boldsymbol{h}$, via the following transformation:

$$\boldsymbol{h}' = \frac{\boldsymbol{h}' - \min(\boldsymbol{h}')}{\max(\boldsymbol{h}') - \min(\boldsymbol{h}')} * \max(\boldsymbol{h}) \quad (11)$$

where $\max()$ and $\min()$ return the maximum and minimum values of an input matrix, respectively. In our experimental analysis, it is validated that this distribution modulation further improves the performance of our coordinate decoding method (Table 3), with the resulting visual effect and qualitative evaluation demonstrated in Fig. 3(b).

Summary We summarise our coordinate decoding method in Fig. 2. Specifically, a total of three steps are involved in a sequence: **(a)** Heatmap distribution modulation (Eq. (10),

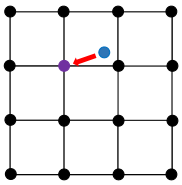

Figure 4: Illustration of quantisation error in the standard coordinate encoding process. The blue point denotes the accurate position ($\boldsymbol{g}'$) of a joint. With the *floor* based coordinate quantisation, an error (indicated by red arrow) is introduced. Other quantisation methods share the same problem.

(11)), **(b)** Distribution-aware joint localisation by Taylor expansion at sub-pixel accuracy (Eq. (3)-(9)), **(c)** Resolution recovery to the original coordinate space (Eq. (2)). None of these steps incur high computational costs, therefore being able to serve as an efficient plug-in for existing models.

Coordinate Encoding

The previous section has addressed the problem with coordinate decoding, rooted at resolution reduction. As a similar process, coordinate encoding shares the same limitation. Specifically, the standard coordinate encoding method starts with downsampling original person images into the model input size. So, the ground-truth joint coordinates require to be transformed accordingly before generating the heatmaps.

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

$$\boldsymbol{g}' = (u', v') = \frac{\boldsymbol{g}}{\lambda} = (\frac{u}{\lambda}, \frac{v}{\lambda}) \tag{12}$$

where λ is the downsampling ratio.

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

$$\boldsymbol{g}'' = (u'', v'') = \text{quantise}(\boldsymbol{g}') = \text{quantise}(\frac{u}{\lambda}, \frac{v}{\lambda}) \tag{13}$$

where quantise() specifies a quantisation function, with the common choices including floor, ceil and round.

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

$$\mathcal{G}(x, y; \boldsymbol{g}'') = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(x-u'')^2 + (y-v'')^2}{2\sigma^2}\right) \tag{14}$$

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

Obviously, the heatmaps generated in the above way are *inaccurate* and *biased* due to the quantisation error (Fig. 4). 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 $\boldsymbol{g}'$ which represents the *accurate* ground-truth coordinate. We still apply Eq. (14) but replacing $\boldsymbol{g}''$ with $\boldsymbol{g}'$. We will demonstrate the benefits of this *unbiased* heatmap generation method (Table 3).

Table 1: Effect of coordinate decoding on the COCO validation set. Model: HRNet-W32; Input size: 128×96.

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	**88.7**	76.3	64.6	72.3	73.7
Ours	**68.4**	88.6	**77.4**	**66.0**	**74.0**	**74.9**

Integration with State-of-the-Art Models

Our DARK method is model-agnostic, seamlessly integrable with any existing heatmap based pose models. Importantly, this does not involve any algorithmic changes to previous methods. In particular, during training the only change is the ground-truth heatmap data generated based on the accurate joint coordinates. At test time, we take as input the predicted heatmaps predicted by any model such as HRNet (Sun et al. 2019), and output more accurate joint coordinates in the original image space. In the whole lifecycle, we keep an existing model intact as the original design. This allows to maximise the generality and scalability of our method.

Experiments

Datasets We used two popular human pose estimation datasets, COCO and MPII. The ***COCO*** keypoint dataset (Lin et al. 2014) presents naturally challenging imagery data with various human poses, unconstrained environments, different body scales and occlusion patterns. The entire objective involves both detecting person instances and localising the body joints. It contains 200,000 images and 250,000 person samples. Each person instance is labelled with 17 joints. The annotations of training and validation sets are publicly benchmarked. In evaluation, we followed the commonly used train2017/val2017/test-dev2017 split. The ***MPII*** human pose dataset (Andriluka et al. 2014) contains 40k person samples, each labelled with 16 joints. We followed the standard train/val/test split as in (Tompson et al. 2014).

Evaluation metrics We used Object Keypoint Similarity (OKS) for COCO and Percentage of Correct Keypoints (PCK) for MPII to evaluate the model performance.

Implementation details For model training, we used the Adam optimiser. For HRNet (Sun et al. 2019) and SimpleBaseline (Xiao, Wu, and Wei 2018), we followed the same learning schedule and epochs as in the original works. For Hourglass (Newell, Yang, and Deng 2016), the base learning rate was fine-tuned to 2.5e-4, and decayed to 2.5e-5 and 2.5e-6 at the 90-th and 120-th epoch. The total number of epochs is 140. We used three different input sizes (128×96, 256×192, 384×288) in our experiments. We adopted the same data preprocessing as in (Sun et al. 2019).

Evaluating Coordinate Representation

As the core problem in this work, the effect of coordinate representation on model performance was firstly examined, with a connection to the input image resolution (size). In this test, by default we used HRNet-W32 (Sun et al. 2019) as the

Table 2: Effect of distribution modulation (DM) on the COCO val set. Backbone: HRNet-W32; Input size: 128×96.

DM	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
✗	68.1	88.5	77.1	65.8	73.7	74.8
✓	**68.4**	**88.6**	**77.4**	**66.0**	**74.0**	**74.9**

Table 3: Effect of coordinate encoding on the COCO validation set. Model: HRNet-W32; Input size: 128 × 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
Unbiased	Standard	**68.0**	**88.9**	**77.0**	**65.4**	**73.7**	**74.5**
Biased	**Ours**	68.4	88.6	77.4	66.0	74.0	74.9
Unbiased	**Ours**	**70.7**	**88.9**	**78.4**	**67.9**	**76.6**	**76.7**

backbone model and 128×96 as the input size, and reported the accuracy results on the COCO validation set.

(i) Coordinate decoding We evaluated the effect of coordinate decoding, in particular, the shifting operation and distribution modulation. The conventional biased heatmaps were used. In this test, we compared the proposed distribution-aware shifting method with *no shifting* (*i.e.* directly using the maximal activation location), and the *standard shifting* (Eq. (1)). We make two major observations in Table 1: **(i)** The standard shifting gives as high as 5.7% AP accuracy boost, which is surprisingly effective. To our best knowledge, this is the first reported effectiveness analysis in the literature, since this problem is largely ignored by previous studies. 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%, among which the distribution modulation gives 0.3% as shown in Table 2. This validates the superiority of our decoding method.

(ii) Coordinate encoding We tested how effective coordinate encoding can be. We compared the proposed *unbiased* encoding with the standard *biased* encoding, along with both the standard and our decoding method. We observed from Table 3 that our unbiased encoding with accurate kernel centre brings positive performance margin, regardless of the coordinate decoding method. In particular, unbiased encoding contributes consistently over 1% AP gain in both cases.

Table 4: Effect of input image size on the COCO validation set. DARK uses HRNet-W32 (HRN32) as backbone.

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

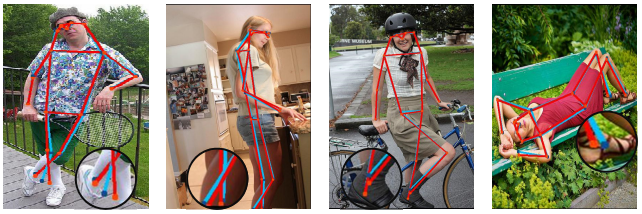

Figure 5: Qualitative evaluation of DARK (red) *vs.* HRNet-W32 (cyan) on COCO.

This suggests the importance of coordinate encoding, which again is neglected by previous investigations.

(iii) Input resolution We examined the impact of input image resolution/size by testing a number of different sizes, considering that it is an important factor relevant to model inference efficiency. 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 4 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). This facilitates the deployment of human pose estimation models on low-resource devices, highly desired in the emerging embedded AI.

(iv) Generality Besides the state-of-the-art HRNet, we also tested other two representative human pose estimation models under varying CNN architectures: SimpleBaseline (Xiao, Wu, and Wei 2018) and Hourglass (Newell, Yang, and Deng 2016). The results in Table 5 show that DARK provides significant performance gain to the existing models in most cases. This suggests a generic usefulness of our approach. We showed qualitative evaluation in Fig. 5.

(v) Complexity We tested the inference efficiency impact by our method in HRNet-W32 at input size of 128×96. On a Titan V GPU, the running speed is reduced from 360 fps to 320 fps in the *low-efficient* python environment, *i.e.* a drop of 11%. We consider this extra cost is rather affordable.

Comparison to the State-of-the-Art Methods

(i) Evaluation on COCO We compared our DARK method with top-performers including G-RMI (Papandreou et al. 2017), Integral Pose Regression (Sun et al. 2018), CPN (Chen et al. 2018), RMPE (Fang et al. 2017), SimpleBaseline (Xiao, Wu, and Wei 2018), and HRNet (Sun et al. 2019). Table 6 shows the accuracy results of the state-of-the-art methods and DARK on the COCO test-dev set. In this test, we used the person detection results from (Sun et al. 2019). We have the following observations: **(i)** DARK with HRNet-W48 at the input size of 384×288 achieves the best accuracy, without extra model parameters and only tiny cost increase. Specifically, compared with the best competitor (HRNet-W48 with the same input size), DARK further improves AP by 0.7% (76.2-75.5). When compared to the most

Table 5: Evaluating the generality of our DARK method to varying state-of-the-art models on the COCO validation set.

DARK	Baseline	Input size	#Params	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
✗	Hourglass (4 Blocks)	128×96	13.0M	2.7	66.2	87.6	75.1	63.8	71.4	72.8
✓					**69.6**	**87.8**	**77.0**	**67.0**	**75.4**	**75.7**
✗	Hourglass (8 Blocks)	128×96	25.1M	4.9	67.6	**88.3**	77.4	65.2	73.0	74.0
✓					**70.8**	87.9	**78.3**	**68.3**	**76.4**	**76.6**
✗	SimpleBaseline-R50	128×96	34.0M	2.3	59.3	85.5	67.4	57.8	63.8	66.6
✓					**62.6**	**86.1**	**70.4**	**60.4**	**67.9**	**69.5**
✗	SimpleBaseline-R101	128×96	53.0M	3.1	58.8	85.3	66.1	57.3	63.4	66.1
✓					**63.2**	**86.2**	**71.1**	**61.2**	**68.5**	**70.0**
✗	SimpleBaseline-R152	128×96	68.6M	3.9	60.7	86.0	69.6	59.0	65.4	68.0
✓					**63.1**	**86.2**	**71.6**	**61.3**	**68.1**	**70.0**
✗	HRNet-W32	128×96	28.5M	1.8	66.9	88.7	76.3	64.6	72.3	73.7
✓					**70.7**	**88.9**	**78.4**	**67.9**	**76.6**	**76.7**
✗	HRNet-W48	128×96	63.6M	3.6	68.0	88.9	77.4	65.7	73.7	74.7
✓					**71.9**	**89.1**	**79.6**	**69.2**	**78.0**	**77.9**

Table 6: Comparison with the state-of-the-art human pose estimation methods on the COCO test-dev set.

Method	Backbone	Input size	#Params	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
G-RMI	ResNet-101	353×257	42.6M	57.0	64.9	85.5	71.3	62.3	70.0	69.7
Integral Pose Regression	ResNet-101	256×256	45.1M	11.0	67.8	88.2	74.8	63.9	74.0	-
CPN	ResNet-Inception	384×288	-	-	72.1	91.4	80.0	68.7	77.2	78.5
RMPE	PyraNet	320×256	28.1M	26.7	72.3	89.2	79.1	68.0	78.6	-
CFN	-	-	-	-	72.6	86.1	69.7	78.3	64.1	-
CPN (ensemble)	ResNet-Inception	384×288	-	-	73.0	91.7	80.9	69.5	78.1	79.0
SimpleBaseline	ResNet-152	384×288	68.6M	35.6	73.7	91.9	81.1	70.3	80.0	79.0
HRNet	HRNet-W32	384×288	28.5M	16.0	74.9	**92.5**	82.8	71.3	80.9	80.1
HRNet	HRNet-W48	384×288	63.6M	32.9	75.5	**92.5**	83.3	71.9	81.5	80.5
DARK	HRNet-W32	128×96	28.5M	**1.8**	70.0	90.9	78.5	67.4	75.0	75.9
DARK	HRNet-W48	384×288	63.6M	32.9	**76.2**	**92.5**	**83.6**	**72.5**	**82.4**	**81.1**
G-RMI (extra data)	ResNet-101	353×257	42.6M	57.0	68.5	87.1	75.5	65.8	73.3	73.3
HRNet (extra data)	HRNet-W48	384×288	63.6M	32.9	77.0	**92.7**	84.5	73.4	83.1	82.0
DARK (extra data)	HRNet-W48	384×288	63.6M	32.9	**77.4**	92.6	**84.6**	**73.6**	**83.7**	**82.3**

Table 7: Comparison on the MPII validation set. DARK uses HRNet-W32 (HRN32) as backbone. Input size: 256×256. Single-scale model performance is considered.

Method	Head	Sho.	Elb.	Wri.	Hip	Kne.	Ank.	Mean
PCKh@0.5								
HRN32	97.1	**95.9**	90.3	86.5	89.1	**87.1**	83.3	90.3
DARK	**97.2**	**95.9**	**91.2**	**86.7**	**89.7**	86.7	**84.0**	**90.6**
PCKh@0.1								
HRN32	51.1	42.7	42.0	41.6	17.9	29.9	31.0	37.7
DARK	**55.2**	**47.8**	**47.4**	**45.2**	**20.1**	**33.4**	**35.4**	**42.0**

efficient model (Integral Pose Regression), DARK(HRNet-W32) achieves an AP gain of 2.2% (70.0-67.8) whilst only needing 16.4% (1.8/11.0 GFLOPs) execution cost. These suggest the advantages and flexibility of DARK on top of existing models in terms of both accuracy and efficiency.

(ii) Evaluation on MPII We compared DARK with HRNet-W32 on the MPII validation set. The comparisons in Table 7 show a consistent performance superiority of our method over the best competitor. Under the more strict accuracy measurement PCKh@0.1, the performance margin of DARK is even more significant. Note, MPII provides significantly smaller training data than COCO, suggesting that our method generalises across varying training data sizes.

Conclusion

In this work, we for the first time systematically investigated the largely ignored yet significant problem of *coordinate representation* (including encoding and decoding) for human pose estimation in unconstrained images. We not only revealed the genuine significance of this problem, but also presented a novel distribution-aware coordinate representation of keypoint (DARK) for more discriminative model

training and inference. Serving as a ready-to-use plug-in component, existing state-of-the-art models can be seamlessly benefited from our DARK method without any algorithmic adaptation at a neglectable cost. Apart from demonstrating empirically the importance of coordinate representation, we validated the performance advantages of DARK by conducting extensive experiments with a wide spectrum of contemporary models on two challenging datasets. We also provided a sequence of in-depth component analysis for giving insights on the design rationale of our model formulation.

References

Andriluka, M.; Pishchulin, L.; Gehler, P.; and Schiele, B. 2014. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Belagiannis, V., and Zisserman, A. 2017. Recurrent human pose estimation. In *IEEE Conference on Automatic Face and Gesture Recognition*.

Carreira, J.; Agrawal, P.; Fragkiadaki, K.; and Malik, J. 2016. Human pose estimation with iterative error feedback. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Chen, X., and Yuille, A. L. 2014. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *Advances in Neural Information Processing Systems*.

Chen, Y.; Shen, C.; Wei, X.-S.; Liu, L.; and Yang, J. 2017. Adversarial posenet: A structure-aware convolutional network for human pose estimation. In *IEEE International Conference on Computer Vision*.

Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; and Sun, J. 2018. Cascaded pyramid network for multi-person pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Chu, X.; Ouyang, W.; Li, H.; and Wang, X. 2016a. Structured feature learning for pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Chu, X.; Ouyang, W.; Wang, X.; et al. 2016b. Crf-cnn: Modeling structured information in human pose estimation. In *Advances in Neural Information Processing Systems*, 316–324.

Fan, X.; Zheng, K.; Lin, Y.; and Wang, S. 2015. Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Fang, H.-S.; Xie, S.; Tai, Y.-W.; and Lu, C. 2017. Rmpe: Regional multi-person pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2334–2343.

Gkioxari, G.; Toshev, A.; and Jaitly, N. 2016. Chained predictions using convolutional neural networks. In *European Conference on Computer Vision*.

Goodfellow, I.; Bengio, Y.; and Courville, A. 2016. *Deep learning*. MIT press.

Ke, L.; Chang, M.-C.; Qi, H.; and Lyu, S. 2018. Multi-scale structure-aware network for human pose estimation. In *European Conference on Computer Vision*.

LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P.; et al. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11):2278–2324.

Lifshitz, I.; Fetaya, E.; and Ullman, S. 2016. Human pose estimation using deep consensus voting. *European Conference on Computer Vision*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*.

Liu, B., and Ferrari, V. 2017. Active learning for human pose estimation. In *IEEE International Conference on Computer Vision*, 4363–4372.

Liu, W.; Chen, J.; Li, C.; Qian, C.; Chu, X.; and Hu, X. 2018. A cascaded inception of inception network with attention modulated feature fusion for human pose estimation. In *AAAI Conference on Artificial Intelligence*.

Newell, A.; Yang, K.; and Deng, J. 2016. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*.

Ning, G.; Zhang, Z.; and He, Z. 2017. Knowledge-guided deep fractal neural networks for human pose estimation. *IEEE Transactions on Multimedia* PP(99):1–1.

Papandreou, G.; Zhu, T.; Kanazawa, N.; Toshev, A.; Tompson, J.; Bregler, C.; and Murphy, K. 2017. Towards accurate multi-person pose estimation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 4903–4911.

Su, K.; Yu, D.; Xu, Z.; Geng, X.; and Wang, C. 2019. Multi-person pose estimation with enhanced channel-wise and spatial information. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Sun, X.; Shang, J.; Liang, S.; and Wei, Y. 2017. Compositional human pose regression. In *IEEE International Conference on Computer Vision*.

Sun, X.; Xiao, B.; Wei, F.; Liang, S.; and Wei, Y. 2018. Integral human pose regression. In *European Conference on Computer Vision*.

Sun, K.; Xiao, B.; Liu, D.; and Wang, J. 2019. Deep high-resolution representation learning for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2818–2826.

Tang, W.; Yu, P.; and Wu, Y. 2018. Deeply learned compositional models for human pose estimation. In *European Conference on Computer Vision*.

Tompson, J. J.; Jain, A.; LeCun, Y.; and Bregler, C. 2014. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in Neural Information Processing Systems*.

Toshev, A., and Szegedy, C. 2014. Deeppose: Human pose estimation via deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Wei, S.-E.; Ramakrishna, V.; Kanade, T.; and Sheikh, Y. 2016. Convolutional pose machines. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Xiao, B.; Wu, H.; and Wei, Y. 2018. Simple baselines for human pose estimation and tracking. In *European Conference on Computer Vision*.

Yang, W.; Ouyang, W.; Li, H.; and Wang, X. 2016. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Yang, W.; Li, S.; Ouyang, W.; Li, H.; and Wang, X. 2017. Learning feature pyramids for human pose estimation. In *IEEE International Conference on Computer Vision*.