24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

Belief Map Enhancement Network for Accurate Human Pose Estimation

Jie Liu[1], Yishun Dou[1], Wenjie Zhang, Jie Tang, Gangshan Wu [2]

Abstract. It is a common practice for pose estimation models to output fixed-size low-resolution belief maps for the body keypoints. The coordinates of the highest belief location are then extracted for each of the body keypoints. When mapping this coarse-grained coordinates back into the fine-grained input space, a minor deviation from the ground-truth location will be magnified many times. So, we can usually get more accurate estimation by using larger belief maps. However, the problem is that we can not use too large belief maps due to the limited computational resources. To alleviate this problem, we propose the Belief Map Enhancement Network (EnhanceNet) for more accurate human pose estimation. EnhanceNet enlarges the belief maps by using the efficient sub-pixel operations, which not only increases the belief map resolution but also corrects some wrong predictions at the same time. Our EnhanceNet is simple yet effective. Extensive experiments are conducted on MPII and COCO datasets to verify the effectiveness of our proposed network. Specifically, we achieve consistently improvements on MPII dataset and COCO human pose dataset by applying our EnhanceNet to the state-of-the-art methods. Our EnhanceNet can be easily inserted into existing networks.

1 Introduction

Human pose estimation refers to the task of precisely localizing important keypoints of human bodies, which serves as an essential technique for a variety of high level tasks, such as activity recognition, tracking and human-computer interaction. It is challenging to achieve accurate localizations due to many confounding factors like pose variation, occlusion and the simultaneous presence of multiple interacting people.

Recently, significant progress on human pose estimation has been made by deep convolutional neural networks (CNNs) [44, 27, 3, 26, 21, 29, 30, 11, 28, 38]. Almost all the CNN based models first down-sample the input image $\mathbf{I}$ to a low-resolution input $\mathbf{I}^{LR}$ very quickly in order to leverage the deep CNN structure to extract high semantic information. To get precise locations, most methods choose to output a belief map $\mathbf{M}^{LR}$ for each body keypoint at the end of networks. To the best of our knowledge, there exists $Size(\mathbf{I}) > Size(\mathbf{I}^{LR}) \geq Size(\mathbf{M}^{LR})$ in all the state-of-the-art methods, where $Size(\cdot)$ represents the spatial resolution. Usually, we first extract intermediate coordinates from $\mathbf{M}^{LR}$ and then map this intermediate coordinates back into input coordinate space by multiplying a factor of value $Size(\mathbf{I})/Size(\mathbf{M}^{LR})$.

[1] Equal contribution

[2] State Key Laboratory for Novel Software Technology. Department of Computer Science and Technology, Nanjing University, China, email: jieliu@smail.nju.edu.cn, tangjie@nju.edu.cn. The corresponding author is Jie Tang.

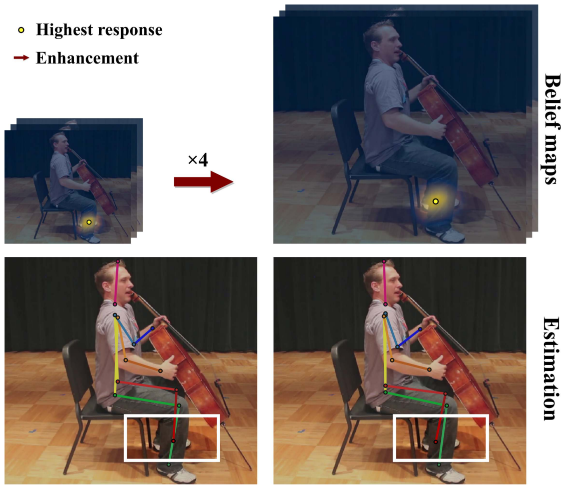

Figure 1: The effects of enhancement. Top-left: original belief maps generated by base models. Top-right: belief maps enhanced by our EnhanceNet. Bottom-left: pose estimation results using original belief maps. Bottom-right: pose estimation results using enhanced belief maps. Our EnhanceNet can make the prediction of left ankle more accurate.

The mapping process can magnify a minor deviation of a predicted body joint many times. As a result, many methods tend to adopt a relatively larger belief map to generate more accurate predications. However, during the deep feature extraction process, we can not maintain a large feature map size until the end of the network. It is more practical to gradually down-sample the input feature maps and then up-sample the feature maps at the tail of network. Due to the high overhead and increasing difficulty of reconstructing high-resolution feature maps from low-resolution feature maps, state-of-the-art methods [27, 39, 26, 5, 45, 38] only continuously up-sample the feature maps to have the same size as $\mathbf{I}^{LR}$, *i.e.* $Size(\mathbf{M}^{LR}) = Size(\mathbf{I}^{LR})$ (see Fig. 2). So, there still exists a big gap between input patch size and output belief map size.

To get larger belief maps, we propose the belief map enhancement network (EnhanceNet) to directly super-resolve the belief maps to a higher resolution (see Fig. 1). This idea was inspired by the success of sub-pixel [35] upscaling in image super-resolution. We also use the sub-pixel upscaling to enlarge belief maps at the end of EnhanceNet. Notice that the purpose of our EnhanceNet is different from the aforementioned feature map up-sampling process. The feature map up-sampling process aims to generate highly representative features for the interpolated locations, where a large number of fea-

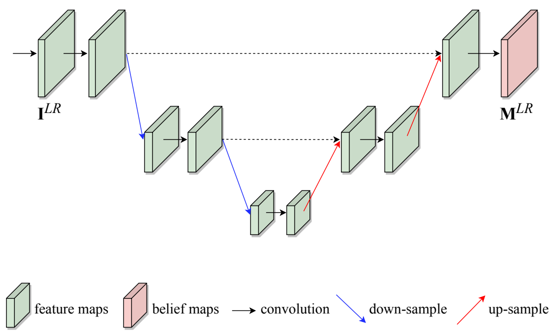

Figure 2: Typical down-sample up-sample process for the trunk of the network. Usually, we have $Size(\mathbf{M}^{LR}) = Size(\mathbf{I}^{LR}) = \frac{1}{4}Size(\mathbf{I})$, where $\mathbf{I}$ is the input image patch. $\mathbf{I}^{LR}$ is down-sampled from $\mathbf{I}$ at the beginning of the network. $\mathbf{M}^{LR}$ is still need to be transformed to the coordinate space of $\mathbf{I}$.

ture channels are needed in order to produce accurate belief maps. In contrast, EnhanceNet uses the belief maps as input and it can enlarge the belief maps by using fewer feature channels without affecting the accuracy. Experimental results show that our proposed EnhanceNet can effectively enhance the belief maps across a wide range of methods on both MPII and COCO datasets.

In summary, our contributions are as follows:

- We propose a belief map enhancement network for highly accurate human pose estimation. Our EnhanceNet can enhance the belief maps with little overhead and obtains much better accuracy.
- We conduct extensive experiments to verify the effectiveness of our EnhanceNet and give a comprehensive analysis of all the details. Our EnhanceNet can consistently improve the performance of state-of-the-art methods on MPII dataset and COCO human pose dataset.

2 RELATED WORK

2.1 Human Pose Estimation

Conventional works on human pose estimation mainly adopt the techniques of pictorial structures [15, 12, 47] or loopy structures [32, 41, 13] to model the spatial relationships of articulated body parts. All of these methods were built on hand-crafted features which are not representative enough to handle severe deformation and occlusion. Recent developments show that earlier methods have been greatly reshaped by convolutional neural networks, which achieve state-of-the-art performance on both single and multi person human pose estimation.

Single Person Pose Estimation. State-of-the-art performance on MPII dataset was mainly achieved by stacked hourglass networks [27] and its follow-ups [46, 8, 20, 39, 48]. Newell *et al.* [27] introduce a novel hourglass module to process and fuse features across multiple scales. They stack up several such hourglass modules, called stacked hourglass networks, to gradually learn long range spatial relationships associated with the body. With the success of stacked hourglass networks, many variants have been proposed. Chu *et al.* [8] incorporate the hourglass module with a multi-context attention mechanism to make the model focus on region of interest. Yang *et al.* [46] design a pyramid residual module to enhance the invariance in scales of the hourglass module. Most recently, some works turn to exploit human skeletally contextual information. Ke *et al.* [20] use structure aware loss to explicitly learn the human skeletal structures. Tang *et al.* [39] further integrate structure supervision into a novel compositional model. Zhang *et al.* [48] introduce a flexible and efficient pose graph neural network to learn a structured representation.

Multi Person Pose Estimation. Multi person pose estimation approaches can be divided into two categories: bottom-up approaches [19, 3, 26, 21, 29] and top-down approaches [30, 11, 16, 5, 45, 38]. Bottom-up approaches directly estimate all keypoints at first and then assemble them into different persons. Part Affinity Field [3] employs a VGG-19 [37] network as a feature encoder, then the output features go through a multi-stage network to produce belief maps and associations of keypoints. Associative Embedding [26] uses the stacked hourglass network to simultaneously output keypoints and group assignments. Top-down approaches firstly locate and crop all persons from the image, and then solve the single person pose estimation task within each patch. Chen *et al.* [5] develop a cascaded pyramid network (CPN) on top of feature pyramid network [22] and propose the online hard keypoints mining (OHKM) strategy. Xiao *et al.* [45] provide a simple yet effective baseline model by appending three stacked deconvolution layers at the end of ResNet [17]. Sun *et al.* [38] propose a novel pose estimation architecture which consists of parallel multi-resolution pathways with repeated information exchange.

2.2 Pose Refinement Networks

Recently, some refinement networks are proposed to refine the estimated poses produced by existing human pose estimation models. Fieraru *et al.* [14] proposed the PoseRefiner that takes as input both the image and a given pose estimate and learns to directly predict a refined pose by jointly reasoning about the input-output space. In order for the network to learn to refine incorrect body keypoint predictions, they employ ad-hoc rules to generate input pose for data augmentation. Similarly, Moon *et al.* [24] proposed the PoseFix refinement network that also takes the estimated pose and original image as input. They used the error statistics as prior information to generate synthetic poses for model training. Different from these pose refinement networks, our EnhanceNet refines the estimated poses by super-resolving the belief maps without any dataset related statistical priors. It takes the belief maps as input and is much more lightweight compared with PoseRefiner and PoseFix.

2.3 Single Image Super-Resolution

Our EnhanceNet is related to single image super-resolution (SR), the task of recovering high-resolution (HR) image from its low-resolution (LR) counterpart. For earlier SR methods, the LR images need to be bicubic interpolated to the desired size before entering the networks, which inevitably increases the computational complexity and might produce new noise. To alleviate this problems, Dong *et al.* [9] exploited the deconvolution operator to upscale spatial resolution at the network tail. Shi *et al.* [35] proposed a more effective sub-pixel convolution layer to replace the deconvolution layer for upscaling the final LR feature maps to HR output. The backbone network for keypoint detection can be seen as a special degradation model that generates LR belief maps, and our EnhanceNet can be seen as a SR model that reconstructs HR ground-truth belief maps from LR belief maps.

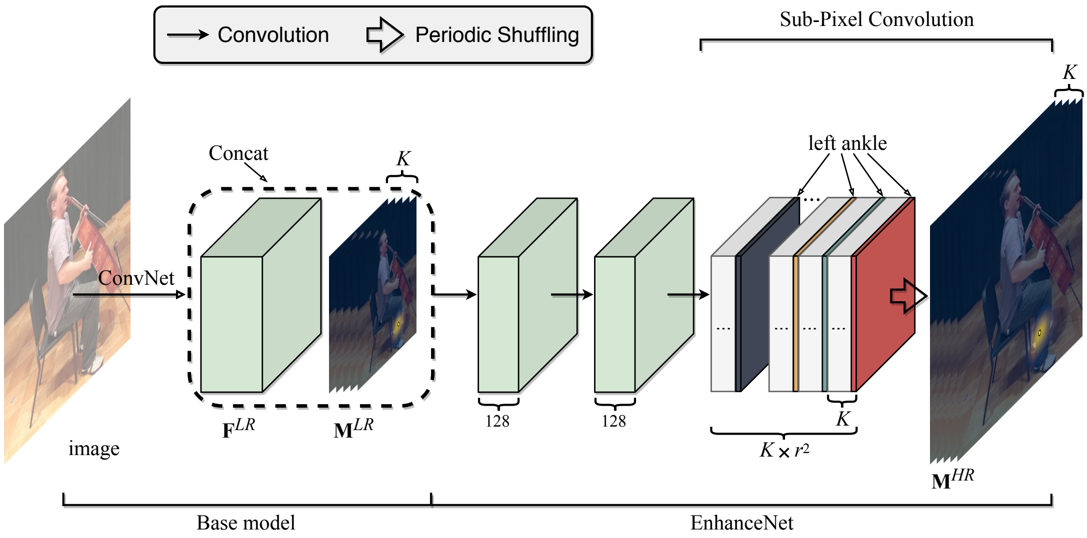

Figure 3: Network architecture of EnhanceNet. $\mathbf{F}^{LR}$ are the feature maps extracted by a pose estimation model (*e.g.* HRNet [38]). $\mathbf{M}^{LR}$ is the low-resolution belief maps. $\mathbf{F}^{LR}$ and $\mathbf{M}^{LR}$ are concatenated as the input of our EnhanceNet which consists of two regular convolution layers and a sub-pixel convolution layer. The sub-pixel convolution layer first generates $K \times r^2$ feature maps, where K is the number of keypoints and r is the upscaling ratio. The final high-resolution belief maps $\mathbf{M}^{HR}$ are then generated by the $\mathcal{PS}$ operation (see Fig. 4).

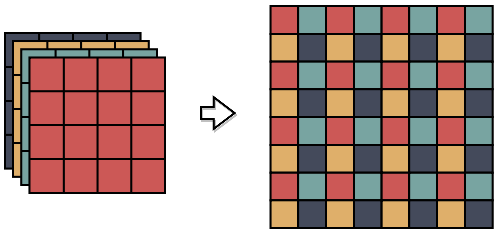

Figure 4: **Periodic Shuffling** ($\mathcal{PS}$) [35] operator in sub-pixel. In this case, the upscaling ratio $r = 2$ and the number of keypoints $K = 1$. The input tensor of size $4 \times 4 \times 2^2$ is rearranged to a tensor of size $8 \times 8 \times 1$.

3 APPROACH

The task of human pose estimation aims to locate body keypoints. Since directly regressing positions [43] from images is a highly non-linear mapping that is difficult to learn, state-of-the-art methods transform this task to estimating belief maps of size $H \times W \times K$ for K body keypoints, where each belief map is a 2D representation of the confidence that a particular body part occurs at each pixel location. In this section, we will describe in detail how the EnhanceNet maps low-resolution belief maps into high-resolution space.

The estimated belief maps of existing models are referred to as $\mathbf{M}^{LR}$ and the super-resolved high-resolution belief maps are referred to as $\mathbf{M}^{HR}$. We denote the last feature maps before generating belief maps of backbone networks as $\mathbf{F}^{LR}$. Both $\mathbf{M}^{LR}$ and $\mathbf{M}^{HR}$ have K channels. The shapes of $\mathbf{M}^{LR}$ and $\mathbf{M}^{HR}$ are $H \times W \times K$ and $rH \times rW \times K$, respectively. Here, r is the upscaling ratio.

3.1 Belief Map Enhancement Network

Conventional pose estimation networks can not continuously increase the feature map resolution to a large scale due to dramatically increased computational cost. Instead, we propose the EnhanceNet to directly enlarge the belief maps generated by pose estimation models, which introduces only a little overhead but achieving much better detection accuracy.

Our EnhanceNet is designed to be simple and effective so that it can be easily inserted into any existing models if applicable. As shown in Fig. 3, we first concatenate $\mathbf{M}^{LR}$ and $\mathbf{F}^{LR}$, then conduct a sequential regular convolution of $L-1$ layers, and finally apply an efficient sub-pixel convolution (the Lth layer) that upscales the low-resolution feature maps to high-resolution belief maps $\mathbf{M}^{HR}$.

For EnhanceNet composed of L layers, the first $L-1$ layers can be described as follows:

$$\mathbf{x} = [\mathbf{F}^{LR}, \mathbf{M}^{LR}] \tag{1}$$

$$f^1(\mathbf{x}) = \mathrm{ReLU}(\mathbf{w}_1^T \mathbf{x}) \tag{2}$$

$$f^l(\mathbf{x}) = \mathrm{ReLU}(\mathbf{w}_l^T f^{l-1}(\mathbf{x})) \tag{3}$$

Biases are absorbed in $\mathbf{w}$ for simplicity. Here $[\cdot,\cdot]$ denotes concatenation and $\mathbf{w}_l$, $l \in \{1, \dots, L-1\}$ are learnable network weights that extract features containing clues for inferring precise locations. The kernel size of $\mathbf{w}_1$ is 1×1 for the purpose of channel reduction, and 3×3 for the rest. The nonlinearity function is ReLU [25].

We adopt sub-pixel [35] convolution layer at the end of the sequential regular convolution layers, where the sub-pixel convolution is an efficient implementation of stride convolution [34] by avoiding convolution happening in high-resolution space. Then $\mathbf{M}^{HR}$ is generated by

$$\mathbf{M}^{HR} = f^L(\mathbf{x}) = \mathcal{PS}(\mathbf{w}_L^T f^{L-1}(\mathbf{x})) \tag{4}$$

Where the weight $\mathbf{w}_L$ has $K \cdot r^2$ filters and $\mathcal{PS}$ [35] is a periodic shuffling operator that rearranges the elements of a $H \times W \times K \cdot r^2$ tensor to a tensor of size $rH \times rW \times K$ without losing information. The effects of this operation is illustrated in Fig. 4. Mathematically, this operation can be described in the following way

$$\mathcal{PS}(\mathbf{T})_{x,y,k} = \mathbf{T}_{\lfloor x/r \rfloor, \lfloor y/r \rfloor, K \cdot r \cdot mod(y,r) + K \cdot mod(x,r) + k} \tag{5}$$

Where (x, y, k) represent coordinates in $\mathbf{M}^{HR}$ of size $rH \times rW \times K$. Notice that the kernel size of $\mathbf{w}_L$ is 3×3, which is greater than commonly used 1×1 since super-resolving high-resolution belief maps needs more contextual information.

3.2 High-resolution Ground-truth and Loss

EnhanceNet is trained together with base models. We use belief maps to represent the body keypoint locations. Denote the ground-truth locations by $\mathbf{z} = \{\mathbf{z}_k\}_{k=1}^{K}$, where $\mathbf{z}_k \in \mathbb{R}^2$ denotes the location of the kth keypoint of a person in the image. Then the high-resolution ground-truth belief map $\mathbf{M}_k^{HR*}$ is generated from a Gaussian with mean $\mathbf{z}_k$ and standard deviation $r\sigma$,

$$\mathbf{M}_k^{HR*}(\mathbf{p}) \sim \mathcal{N}(\mathbf{z}_k, (r\sigma)^2) \tag{6}$$

Where $\mathbf{p} \in \mathbb{R}^2$ denotes the location, and σ is the standard deviation in generating the low-resolution ground-truth belief maps. Notice that bottom-up approaches predict keypoints of different persons simultaneously, where multi-peak ground-truth belief maps are required. When combining multiple belief maps into a single one, we take the maximum of individual belief maps of each person.

EnhanceNet estimates K bilief maps, *i.e.* $\mathbf{M}^{HR} = \{\mathbf{M}_k^{HR}\}_{k=1}^{K}$, for K body keypoints. We adopt Mean Squared Error (MSE) loss for model training. Given N input patches, the loss is defined by

$$\mathcal{L}^{HR} = \sum_{n=1}^{N}\sum_{k=1}^{K} ||\mathbf{M}_k^{HR*} - \mathbf{M}_k^{HR}||^2 \tag{7}$$

Combined with the loss $\mathcal{L}^{LR}$ in base model, the total loss is

$$\mathcal{L} = \mathcal{L}^{LR} + \eta\mathcal{L}^{HR} \tag{8}$$

Where η is the balance factor and we set η to 1 in all of our experiments.

3.3 Sub-pixel *vs*. Deconv *vs*. Interpolation

We adopt sub-pixel convolution as the upscaling layer at the end of our EnhanceNet. Sub-pixel convolution is an essential component in the task of image super-resolution. Deconvolution is also commonly used to increase resolution [10, 31, 34, 18]. However, deconvolution with small kernel size may not perform well at large upscale ratio (*e.g.* $\times 4$), thus a larger kernel size (*e.g.* > 10) is typically used [34, 18].

In fact, as discussed in [36] the effect of a sub-pixel convolution layer with weight shape $(C_{in}, C_{out} \times r^2, k_h, k_w)$ is identical to that of a deconvolution layer with weight shape $(C_{in}, C_{out}, k_h \times r, k_w \times r)$, where C_{in}, C_{out}, k, r represent input channels, output channels, kernel size and upscaling ratio, respectively. In this case, the two have the same number of parameters, represented as P. The spatial resolution $H \times W$ is maintained after sub-pixel convolution, but expanded to $rH \times rW$ after deconvolution. Accordingly, the GFLOPs of one sub-pixel convolution layer is $H \times W \times P$; and $rH \times rW \times P$ for deconvolution, which is r^2 times that of sub-pixel.

Another widely-used way to upscale low-resolution feature maps is interpolation followed by a convolution [22, 4]. Assume that the kernel size is the same with that of sub-pixel convolution, then the weight shape is $(C_{in}, C_{out}, k_h, k_w)$, which may be lack of representation power because the number of parameters is only $1/r^2$ times that of sub-pixel. Meanwhile, the GFLOPs is same with that of sub-pixel convolution since the convolution happens in the upscaled space. Moreover, the receptive field is smaller than sub-pixel convolution and may degrade the performance when applying to our EnhanceNet.

In a word, the sub-pixel convolution is more powerful when having the same computational complexity in the case of our EnhanceNet, which is consistent with our experimental results in Table 5c.

4 Experiments

We verify the effectiveness and generality of EnhanceNet on both single and multi person pose estimation across multiple leading methods. All the models are trained using officially published open source code. All the reported results use the models we re-trained from scratch. There may exist a slight difference between the original paper and that we reported. It does not matter since we mainly concern with the improvement. We set the number of layers $L = 3$, the number of channels $C = 128$ and the upscaling factor $r = 4$ in our EnhanceNet. For single person pose estimation the input patch size is 256×256 and for multi person estimation the input patch size is 256×192 except for Associative Embedding [26] whose patch size is 512×512.

4.1 Single Person Pose Estimation

Dataset. The MPII Human Pose dataset [1] consists of around 25k images with 40k annotated samples (28k for training, 11k for testing), which covers a wide range of real-world activities and a great variety of full-body poses. We evaluate proposed EnhanceNet on the validation set and test set, where the validation set contains 3k samples split from training set following [42, 27]. Different from the recent leading method [48], we do not include any extra training data.

Evaluation Metric. Following previous work, we use the PCKh (head-normalized Percentage of Corrected Keypoints) score as the evaluation metric. A keypoint is correct if it falls within αl pixels from the ground-truth location, where l is the ground-truth head length and α is a threshold that controls the tolerance of jitter errors. The improvement on PCKh@0.5 ($\alpha = 0.5$) score is reported. In addition, we also do comparisons at stricter thresholds (smaller α).

Methods	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
Hourglass (2 stage) [27]	**96.08**	94.74	88.24	82.87	86.91	81.95	78.44	87.14
+EnhanceNet	95.70	**95.14**	**89.13**	**84.00**	**87.35**	**84.12**	**79.38**	**87.96**
Hourglass (4 stage)	96.49	95.50	88.99	84.46	87.43	**84.65**	80.21	88.34
+EnhanceNet	**96.73**	**95.57**	**89.76**	**85.06**	**88.51**	84.42	**81.03**	**88.81**
Hourglass (8 stage)	96.79	95.28	90.27	**85.56**	87.57	84.30	81.06	88.78
+EnhanceNet	**96.79**	**95.41**	**90.30**	85.41	**88.14**	**84.85**	**81.25**	**89.03**
DLCM [39]	96.78	96.03	90.88	**86.96**	89.74	86.90	82.57	90.37
+EnhanceNet	**97.53**	**96.25**	**91.26**	86.89	**90.36**	**86.90**	**83.61**	**90.78**

Table 1: Improvement of PCKh@0.5 when EnhanceNet is applied to the state-of-the-art single person pose estimation methods. The PCKh@0.5 is calculated on the MPII validation set.

Performance improvement. Table 1 shows the improvements of PCKh@0.5 score on the MPII validation set when our EnhanceNet is applied to state-of-the-art single person pose estimation methods, *e.g.* stacked hourglass [27] and DLCM [39], where DLCM achieved 92.3 PCKh@0.5 score and ranked first on MPII leaderboard among the methods without using extra training data. By adding EnhanceNet, the PCKh@0.5 score of DLCM improves from 90.37 to 90.78 on the

Methods	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
Insafutdinov *et al.* [19]	96.8	95.2	89.3	84.4	88.4	83.4	78.0	88.5
Wei *et al.* [44]	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Bulat *et al.* [2]	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7
Newell *et al.* [27]	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9
Tang *et al.* [40]	97.4	96.4	92.1	87.7	90.2	87.7	84.3	91.2
Chu *et al.* [8]	98.5	96.3	91.9	88.1	90.6	88.0	85.0	91.5
Chou *et al.* [7]	98.2	96.8	92.2	88.0	91.3	89.1	84.9	91.8
Chen *et al.* [6]	98.1	96.5	92.5	88.5	90.2	89.6	86.0	91.9
Yang *et al.* [46]	98.5	96.7	92.5	88.7	91.1	88.6	86.0	92.0
Ke *et al.* [20]	98.5	96.8	92.7	88.4	90.6	89.4	86.3	92.1
SimpleBaseline [45]	98.5	96.6	91.9	87.6	91.1	88.1	84.1	91.5
HRNet-W32 [38]	98.6	96.9	92.8	**89.0**	91.5	89.0	85.7	92.3
DLCM [39]	98.4	96.9	92.6	88.7	**91.8**	89.4	86.2	92.3
DLCM + EnhanceNet	**98.6**	**97.0**	**92.8**	88.8	91.7	**89.6**	**86.6**	**92.5**

Table 2: Results of PCKh@0.5 on the MPII test set.

validation set. For stacked hourglass network, we achieve consistent improvements with different number of stages.

Table 2 shows the PCKh@0.5 score on the MPII test set. A simple addition of EnhanceNet on DLCM establishes a new state-of–the-art on MPII test set. Notice that HRNet [38] achieves the same performance as DLCM by using HRNet-W32, but the performance is stagnant when they turn to a much bigger network (HRNet-W48) that has double complexity of HRNet-W32 in terms of both parameters and GFLOPs. In contrast, our EnhanceNet causes a new state-of-the-art with only a little overhead. Qualitative results on MPII are presented in Fig. 6a.

Challenging Threshold. It is worthy to note that our EnhanceNet shows even better performance at a more challenging threshold *i.e.* PCKh@0.1. As shown in Table 3, the top-performed DLCM obtains significant improvements by applying our EnhanceNet: 2.55 points gain for mean score, and even 4.1 points gain for head. Furthermore, we compare PCKh score at all thresholds in Fig. 5. DLCM get consistent improvements at all thresholds on both the most accurate (*i.e.* Head) and the most challenging (*i.e.* Ankle) body keypoints. The large improvements at strict thresholds indicate that our EnhanceNet is capable of generating high-resolution belief maps, which is more suitable for high precision keypoints detection.

Methods	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
DLCM [39]	49.74	39.88	39.54	38.89	17.34	27.67	28.84	35.00
+EnhanceNet	**53.84**	**42.48**	**43.43**	**40.67**	**18.40**	**29.34**	**31.44**	**37.55**

Table 3: Improvement of PCKh@0.1 when EnhanceNet is applied to DLCM. There is a significant improvement of 2.55 points at this challenging threshold. The PCKh@0.1 is calculated on the MPII validation set.

4.2 Multi Person Pose Estimation

Dataset. The MS COCO dataset [23] contains more than 200k images and 250k person instances labels with keypoints. We train all the models on COCO train2017 set, containing 57k images and 150k person instances. We evaluate proposed EnhanceNet on the val2017 set and test-dev2017 set, including 5k images and 20k images, respectively.

Evaluation Metric. The evaluation defines the object keypoint similarity (OKS) and uses the mean average precision (AP) over 10 OKS thresholds as main competition metric [23]. The OKS plays the same role as the IoU in object detection. It is calculated from scale of person and the distance between predicted points and ground-truth

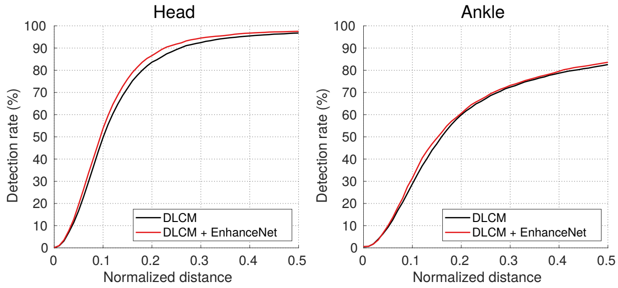

Figure 5: Comparisons of PCKh curves on head and ankle when EnhanceNet is applied to DLCM [39]. The improvement of PCKh@0.1 is remarkable: 4.1 points for head and 2.6 points for ankle, indicating a strong localization performance improvement. The PCKh score is calculated on the MPII validation set.

points. We report standard average precisions and recall scores: AP, $AP_{oks=0.50}$, $AP_{oks=0.75}$, $AP_{Medium\ obj}$, $AP_{Large\ obj}$ and AR.

Testing. Top-down methods adopt a two-stage paradigm: detect the persons using a detector and estimate keypoints locations. For person detection, we use detection results provided by SimpleBaseline [45] with person category AP 56.4 on val2017 set, and 60.9 on test-dev2017.

Methods	AP	$AP_{.50}$	$AP_{.75}$	AP_M	AP_L	AR
Associative Embedding [26]	53.3	77.0	57.7	43.3	69.1	60.7
+EnhanceNet	**55.3**	**77.9**	**60.4**	**45.4**	**70.7**	**62.5**
CPN (ResNet-50) [5]	69.1	**87.7**	76.2	65.7	76.0	76.5
+EnhanceNet	**70.0**	87.5	**76.9**	**67.2**	**76.5**	**77.7**
CPN (ResNet-101)	69.7	87.7	76.9	66.6	76.3	77.0
+EnhanceNet	**70.5**	**87.7**	**77.6**	**67.7**	**76.7**	**78.0**
SimpleBaseline (ResNet-50) [45]	70.2	88.7	77.7	67.0	76.9	76.1
+EnhanceNet	**71.3**	**88.9**	**78.5**	**68.0**	**77.9**	**77.1**
SimpleBaseline (ResNet-101)	71.4	89.2	79.1	68.1	78.2	77.2
+EnhanceNet	**72.1**	**89.2**	**79.4**	**68.6**	**79.0**	**77.8**
HRNet (HRNet-W32) [38]	74.3	89.9	81.6	70.8	81.0	79.7
+EnhanceNet	**75.1**	**90.4**	**82.1**	**71.6**	**81.6**	**80.3**
HRNet (HRNet-W48)	75.0	90.4	82.2	71.3	82.1	80.4
+EnhanceNet	**75.8**	**90.6**	**82.5**	**72.1**	**82.7**	**81.0**

Table 4: Improvement of APs when EnhanceNet is applied to the state-of-the-art multi person pose estimation methods. The APs are calculated on the COCO val2017 set.

Performance Improvement. Table 4 and Table 6 show the improvements on val2017 and test-dev2017 sets when our EnhanceNet is applied to state-of-the-art multi person pose estimation methods: Associative Embedding [26], CPN [5], SimpleBaseline [45] and HRNet [38]. For Associative Embedding, we keep the embedding branch intact and add our EnhanceNet to the detection branch to enhance the belief maps for detected keypoints. By adding EnhanceNet, the AP of associative embedding improved by around 2 points on both val2017 and test-dev2017 sets. CPN adopts online hard keypoints mining (OHKM) that only punish the losses of hard keypoints. We put the OHKM at the end of our EnhanceNet and achieve about 1 point improvement on both ResNet-50 and ResNet-101. SimpleBaseline consists of a ResNet and three stacked deconvolution layers. We directly append EnhanceNet at the end of SimpleBaseline and obtain consistent improvements on val2017 and test-dev2017 sets. HRNet, the top-performed method on COCO human pose leaderboard, also gains considerable improvements: 0.8 points for HRNet-W32 and its wider counterpart HRNet-W48.

(a) MPII

(b) MS COCO

Figure 6: Qualitative results on the MPII validation set and the COCO val2017 set, before and after applying the proposed EnhanceNet on top-performed methods. The white rectangles denote the areas where EnhanceNet brings significant improvement. Our enhancement method provides better localization, it can relieve small displacement error and predict highly precise positions (Best viewed in electronic form with 4× zoom in).

	AP	$AP_{.50}$	$AP_{.75}$	AR
$\mathbf{F}^{LR}$	71.0	88.7	78.3	76.7
$\mathbf{M}^{LR}$	71.1	88.9	78.3	77.0
$\mathbf{M}^{LR}+\mathbf{F}^{LR}$	**71.3**	**88.9**	**78.5**	**77.1**

(a) **Input of EnhanceNet**: Decomposing the input of EnhanceNet. $\mathbf{M}^{LR}$ and $\mathbf{F}^{LR}$ represent low-resolution belief maps and low-resolution feature maps, respectively.

	C_{in}	C_{out}	kernel size	#Params	GFLOPs	AP	$AP_{.50}$	$AP_{.75}$	AR
bilinear + conv	128	17	3×3	0.20M	1.52	70.9	88.9	78.2	76.8
deconv	128	17	3×3	0.20M	1.52	70.8	88.8	78.1	76.6
deconv	128	17	12×12	0.50M	15.96	71.2	88.7	78.5	76.9
sub-pixel	128	272	3×3	0.50M	1.52	**71.3**	**88.9**	**78.5**	**77.1**

(c) **Upscaling layer**: Performance comparisons of EnhanceNet with different upscaling layers. C_{in} and C_{out} represent the number of input channel and output channel, respectively.

Channels	#Params	GFLOPs	AP	$AP_{.50}$	$AP_{.75}$	AR
$C=64$	0.21M	0.65	70.8	88.7	78.2	76.7
$C=128$	0.50M	1.52	71.3	88.9	78.5	77.1
$C=256$	1.29M	3.95	71.4	89.0	78.5	77.2

(b) **Channels** C: Performance comparisons with different number of channels in our EnhanceNet.

	#Params	GFLOPs	AP	$AP_{.50}$	$AP_{.75}$	AR
base-model	-	-	70.2	88.7	77.7	76.1
$r=2$	0.26M	0.80	70.4	88.8	78.4	76.3
$r=3$	0.36M	1.10	71.0	88.6	78.5	76.7
$r=4$	0.50M	1.52	**71.3**	**88.9**	78.5	**77.1**
$r=5$	0.67M	2.06	71.2	88.8	**78.6**	77.0

(d) **Upscaling ratio** (r): The performance of EnhanceNet at various upscaling ratios.

Table 5: **Ablations** on COCO keypoints detection when the EnhanceNet is applied to SimpleBaseline (ResNet-50) [45]. The #Params and GFLOPs are calculated on our EnhanceNet. The AP and AR scores are calculated on val2017 set.

Methods	AP	$AP_{.50}$	$AP_{.75}$	AP_M	AP_L	AR
Associative Embedding [26]	54.6	80.4	59.1	44.9	68.3	60.3
+EnhanceNet	**56.9**	**80.7**	**61.8**	**47.6**	**70.1**	**62.6**
CPN (ResNet-50) [5]	68.7	89.5	76.6	65.7	74.2	75.7
+EnhanceNet	**69.8**	**89.8**	**77.5**	**67.0**	**75.3**	**77.1**
CPN (ResNet-101)	69.1	89.7	77.2	66.2	74.7	76.3
+EnhanceNet	**70.0**	**90.0**	**77.9**	**67.4**	**75.4**	**77.4**
SimpleBaseline (ResNet-50) [45]	69.8	90.8	77.9	66.6	75.6	75.5
+EnhanceNet	**70.9**	**91.0**	**78.8**	**67.6**	**76.8**	**76.4**
SimpleBaseline (ResNet-101)	70.7	91.1	79.2	67.8	76.3	76.5
+EnhanceNet	**71.6**	**91.1**	**79.8**	**68.5**	**77.4**	**77.2**
HRNet (HRNet-W32) [38]	73.4	92.1	81.7	70.2	79.3	78.9
+EnhanceNet	**74.2**	**92.1**	**82.1**	**70.9**	**79.9**	**79.4**
HRNet (HRNet-W48)	74.1	92.3	82.2	70.8	79.8	79.5
+EnhanceNet	**74.9**	**92.3**	**82.8**	**71.6**	**80.6**	**80.1**

Table 6: Improvement of APs when EnhanceNet is applied to the state-of-the-art multi person pose estimation methods. The APs are calculated on the COCO test-dev2017 set.

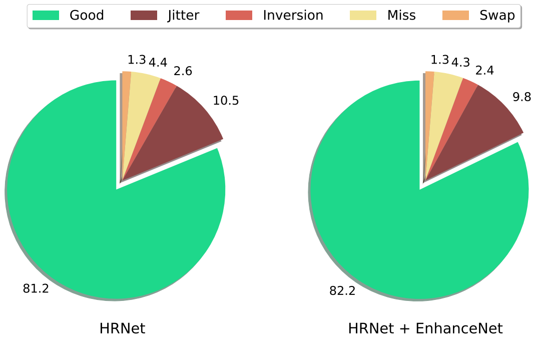

Figure 7: Frequency changes of each error type when the EnhanceNet is applied to HRNet-W32 [38]. The frequency is calculated on the COCO val2017 set.

The improvement brought by our EnhanceNet is not only because it increases the depth of base model. To see this, we note that CPN with ResNet-50 has 70.0 and 69.8 AP on val2017 and test-dev2017 sets when adding our EnhanceNet. However, the original CPN with ResNet-101 has only 69.7 and 69.1 AP, respectively. Similar phenomenon can also be found in SimpleBaseline and HRNet. This indicates that our EnhanceNet can effectively enhance the belief maps generated by base models and is able to predict more precise keypoint locations. Qualitative results on COCO are presented in Fig. 6b.

Error Frequency Change. To better understand the behavior of EnhanceNet and find out how it improves the performance, we analyze the frequency changes of each error type when it is applied to HRNet-W32. As shown in Fig. 7, the gains mainly come from the correction of small displacement error (*i.e. jitter*) [33], which further proves the effectiveness of our EnhanceNet.

5 Ablation Study

In this section, we provide an in-depth analysis of each individual design of our EnhanceNet. All the experiments in Table 5 are conducted on SimpleBaseline [45] with ResNet-50.

Input of EnhanceNet. In Table 5a we study the effects of different inputs fed into EnhanceNet. A competitive result can be obtained even if only the low-resolution belief maps (*i.e.* $\mathbf{M}^{LR}$) are used as input. This indicates that our EnhanceNet can truly enhance the belief maps by only super-resolving them. Interestingly, only using the low-resolution feature maps $\mathbf{F}^{LR}$ can not bring more improvement than using $\mathbf{M}^{LR}$, this indicates that our EnhanceNet mainly gains improvement by enhancing the belief maps. The best performance is achieved by concatenating $\mathbf{M}^{LR}$ and $\mathbf{F}^{LR}$, where the $\mathbf{F}^{LR}$ contains rich semantic information which is helpful for enhancing the belief maps.

Number of Channels. Table 5b shows the performance comparisons of our EnhanceNet with different number of feature channels. The performance at $C = 128$ is almost as good as $C = 256$, which indicates that our EnhanceNet can behave well with a low computational cost.

Upscaling Layer. We compare the complexity and performance of different upscaling layers in Table 5c. The interpolation is instantiated with bilinear and the convolution has same kernel size with that of sub-pixel. The interpolation combined with convolution has same computational complexity with sub-pixel but achieves a lower AP. As discussed in § 3.3, deconvolution can have the same effect as sub-pixel convolution when using a large kernel. This is consistent with our experiments: when the kernel size is 12, deconvolution achieves similar AP with sub-pixel convolution; but when using a kernel size of 3, the AP dropped by 0.4 which is unacceptable. However, when using a kernel size of 12, the GFLOPs is much higher than that of sub-pixel convolution. So, we can conclude that sub-pixel convolution is most suitable for our belief map enhancement network.

Upscaling Ratio. Table 5d shows the performance of EnhanceNet at various upscaling ratios. The number of parameters and GFLOPs are only counted for our EnhanceNet. As we can see, the best performance is achieved at $r = 4$. When $r = 5$, the model complexity increases but the detection performance has not been improved accordingly. We choose $r = 4$ as the upscaling ratio of EnhanceNet since it has the best trade-off between model complexity and keypoint detection performance.

6 Conclusion

In this paper, we proposed a belief map enhancement network (EnhanceNet) to enlarge the belief maps generated by existing human pose models and correct some wrong predictions at the same time. Our EnhanceNet can be easily inserted into state-of-the-art pose estimation models. By using EnhanceNet, we achieve consistently improvements on MPII dataset and COCO human pose dataset across multiple leading methods. Extensive experiments have shown the effectiveness of our EnhanceNet.

REFERENCES

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter V. Gehler, and Bernt Schiele, '2d human pose estimation: New benchmark and state of the art analysis', in *CVPR*, pp. 3686–3693. IEEE Computer Society, (2014).
[2] Adrian Bulat and Georgios Tzimiropoulos, 'Human pose estimation via convolutional part heatmap regression', in *ECCV (7)*, volume 9911 of *Lecture Notes in Computer Science*, pp. 717–732. Springer, (2016).
[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh, 'Realtime multi-person 2d pose estimation using part affinity fields', in *CVPR*, pp. 1302–1310. IEEE Computer Society, (2017).
[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam, 'Encoder-decoder with atrous separable convolution for semantic image segmentation', in *ECCV (7)*, volume 11211 of *Lecture Notes in Computer Science*, pp. 833–851. Springer, (2018).
[5] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun, 'Cascaded pyramid network for multi-person pose estimation', in *CVPR*, pp. 7103–7112. IEEE Computer Society, (2018).

[6] Yu Chen, Chunhua Shen, Xiu-Shen Wei, Lingqiao Liu, and Jian Yang, 'Adversarial posenet: A structure-aware convolutional network for human pose estimation', in *ICCV*, pp. 1221–1230. IEEE Computer Society, (2017).
[7] Chia-Jung Chou, Jui-Ting Chien, and Hwann-Tzong Chen, 'Self adversarial training for human pose estimation', in *APSIPA*, pp. 17–30. IEEE, (2018).
[8] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L. Yuille, and Xiaogang Wang, 'Multi-context attention for human pose estimation', in *CVPR*, pp. 5669–5678. IEEE Computer Society, (2017).
[9] Chao Dong, Chen Change Loy, and Xiaoou Tang, 'Accelerating the super-resolution convolutional neural network', in *ECCV (2)*, volume 9906 of *Lecture Notes in Computer Science*, pp. 391–407. Springer, (2016).
[10] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Häusser, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, and Thomas Brox, 'Flownet: Learning optical flow with convolutional networks', in *ICCV*, pp. 2758–2766. IEEE Computer Society, (2015).
[11] Haoshu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu, 'RMPE: regional multi-person pose estimation', in *ICCV*, pp. 2353–2362. IEEE Computer Society, (2017).
[12] Pedro F. Felzenszwalb and Daniel P. Huttenlocher, 'Pictorial structures for object recognition', *International Journal of Computer Vision*, **61**(1), 55–79, (2005).
[13] Vittorio Ferrari, Manuel J. Marín-Jiménez, and Andrew Zisserman, '2d human pose estimation in TV shows', in *Statistical and Geometrical Approaches to Visual Motion Analysis*, volume 5604 of *Lecture Notes in Computer Science*, pp. 128–147. Springer, (2008).
[14] Mihai Fieraru, Anna Khoreva, Leonid Pishchulin, and Bernt Schiele, 'Learning to refine human pose estimation', in *CVPR Workshops*, pp. 205–214. IEEE Computer Society, (2018).
[15] Martin A. Fischler and Robert A. Elschlager, 'The representation and matching of pictorial structures', *IEEE Trans. Computers*, **22**(1), 67–92, (1973).
[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick, 'Mask R-CNN', in *ICCV*, pp. 2980–2988. IEEE Computer Society, (2017).
[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, 'Deep residual learning for image recognition', in *CVPR*, pp. 770–778. IEEE Computer Society, (2016).
[18] Zheng Hui, Xiumei Wang, and Xinbo Gao, 'Fast and accurate single image super-resolution via information distillation network', in *CVPR*, pp. 723–731. IEEE Computer Society, (2018).
[19] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele, 'Deepercut: A deeper, stronger, and faster multi-person pose estimation model', in *ECCV (6)*, volume 9910 of *Lecture Notes in Computer Science*, pp. 34–50. Springer, (2016).
[20] Lipeng Ke, Ming-Ching Chang, Honggang Qi, and Siwei Lyu, 'Multi-scale structure-aware network for human pose estimation', in *ECCV (2)*, volume 11206 of *Lecture Notes in Computer Science*, pp. 731–746. Springer, (2018).
[21] Muhammed Kocabas, Salih Karagoz, and Emre Akbas, 'Multiposenet: Fast multi-person pose estimation using pose residual network', in *ECCV (11)*, volume 11215 of *Lecture Notes in Computer Science*, pp. 437–453. Springer, (2018).
[22] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie, 'Feature pyramid networks for object detection', in *CVPR*, pp. 936–944. IEEE Computer Society, (2017).
[23] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick, 'Microsoft COCO: common objects in context', in *ECCV (5)*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, (2014).
[24] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee, 'Posefix: Model-agnostic general human pose refinement network', *arXiv preprint*, **arXiv:1812.03595**, (2018).
[25] Vinod Nair and Geoffrey E. Hinton, 'Rectified linear units improve restricted boltzmann machines', in *ICML*, pp. 807–814. Omnipress, (2010).
[26] Alejandro Newell, Zhiao Huang, and Jia Deng, 'Associative embedding: End-to-end learning for joint detection and grouping', in *NIPS*, pp. 2274–2284, (2017).
[27] Alejandro Newell, Kaiyu Yang, and Jia Deng, 'Stacked hourglass networks for human pose estimation', in *ECCV (8)*, volume 9912 of *Lecture Notes in Computer Science*, pp. 483–499. Springer, (2016).
[28] Xuecheng Nie, Jiashi Feng, Junliang Xing, and Shuicheng Yan, 'Pose partition networks for multi-person pose estimation', in *ECCV (5)*, volume 11209 of *Lecture Notes in Computer Science*, pp. 705–720. Springer, (2018).
[29] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy, 'Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model', in *ECCV (14)*, volume 11218 of *Lecture Notes in Computer Science*, pp. 282–299. Springer, (2018).
[30] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy, 'Towards accurate multi-person pose estimation in the wild', in *CVPR*, pp. 3711–3719. IEEE Computer Society, (2017).
[31] Alec Radford, Luke Metz, and Soumith Chintala, 'Unsupervised representation learning with deep convolutional generative adversarial networks', in *ICLR*, (2016).
[32] Xiaofeng Ren, Alexander C. Berg, and Jitendra Malik, 'Recovering human body configurations using pairwise constraints between parts', in *ICCV*, pp. 824–831. IEEE Computer Society, (2005).
[33] Matteo Ruggero Ronchi and Pietro Perona, 'Benchmarking and error diagnosis in multi-instance pose estimation', in *ICCV*, pp. 369–378. IEEE Computer Society, (2017).
[34] Evan Shelhamer, Jonathan Long, and Trevor Darrell, 'Fully convolutional networks for semantic segmentation', *IEEE Trans. Pattern Anal. Mach. Intell.*, **39**(4), 640–651, (2017).
[35] Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang, 'Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network', in *CVPR*, pp. 1874–1883. IEEE Computer Society, (2016).
[36] Wenzhe Shi, Jose Caballero, Lucas Theis, Ferenc Huszar, Andrew P. Aitken, Christian Ledig, and Zehan Wang, 'Is the deconvolution layer the same as a convolutional layer?', *arXiv preprint*, **arXiv:1609.07009**, (2016).
[37] Karen Simonyan and Andrew Zisserman, 'Very deep convolutional networks for large-scale image recognition', in *ICLR*, (2015).
[38] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang, 'Deep high-resolution representation learning for human pose estimation', *arXiv preprint*, **arXiv:1902.09212**, (2019).
[39] Wei Tang, Pei Yu, and Ying Wu, 'Deeply learned compositional models for human pose estimation', in *ECCV (3)*, volume 11207 of *Lecture Notes in Computer Science*, pp. 197–214. Springer, (2018).
[40] Zhiqiang Tang, Xi Peng, Shijie Geng, Lingfei Wu, Shaoting Zhang, and Dimitris N. Metaxas, 'Quantized densely connected u-nets for efficient landmark localization', in *ECCV (3)*, volume 11207 of *Lecture Notes in Computer Science*, pp. 348–364. Springer, (2018).
[41] Tai-Peng Tian and Stan Sclaroff, 'Fast globally optimal 2d human detection with loopy graph models', in *CVPR*, pp. 81–88. IEEE Computer Society, (2010).
[42] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler, 'Efficient object localization using convolutional networks', in *CVPR*, pp. 648–656. IEEE Computer Society, (2015).
[43] Alexander Toshev and Christian Szegedy, 'Deeppose: Human pose estimation via deep neural networks', in *CVPR*, pp. 1653–1660. IEEE Computer Society, (2014).
[44] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh, 'Convolutional pose machines', in *CVPR*, pp. 4724–4732. IEEE Computer Society, (2016).
[45] Bin Xiao, Haiping Wu, and Yichen Wei, 'Simple baselines for human pose estimation and tracking', in *ECCV (6)*, volume 11210 of *Lecture Notes in Computer Science*, pp. 472–487. Springer, (2018).
[46] Wei Yang, Shuang Li, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang, 'Learning feature pyramids for human pose estimation', in *ICCV*, pp. 1290–1299. IEEE Computer Society, (2017).
[47] Yi Yang and Deva Ramanan, 'Articulated pose estimation with flexible mixtures-of-parts', in *CVPR*, pp. 1385–1392. IEEE Computer Society, (2011).
[48] Hong Zhang, Hao Ouyang, Shu Liu, Xiaojuan Qi, Xiaoyong Shen, Ruigang Yang, and Jiaya Jia, 'Human pose estimation with spatial contextual information', *arXiv preprint*, **arXiv:1901.01760**, (2019).