Belief Map Enhancement Network for Accurate Human Pose Estimation

Jie Liu, Yishun Dou, Wenjie Zhang, Jie Tang, Gangshan Wu

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 coordinate 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 \( I \) to a low-resolution input \( I^{LR} \) 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 \( M^{LR} \) for each body keypoint at the end of networks. To the best of our knowledge, there exists \( \text{Size}(I) > \text{Size}(I^{LR}) \geq \text{Size}(M^{LR}) \) in all the state-of-the-art methods, where \( \text{Size}() \) represents the spatial resolution. Usually, we first extract intermediate coordinates from \( M^{LR} \) and then map this intermediate coordinates back into input coordinate space by multiplying a factor of value \( \frac{\text{Size}(I)}{\text{Size}(M^{LR})} \).

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 predictions. 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 \( I^{LR} \), i.e. \( \text{Size}(M^{LR}) = \text{Size}(I^{LR}) \) (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] upsampling 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 upsampling process. The feature map upsampling process aims to generate highly representative features for the interpolated locations, where a large number of fea-
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 skeletonly 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 problem, 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.
Conventional pose estimation networks can not continuously in-
crease the feature map resolution to a large scale due to dramatically
This operation can be described in the following way
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 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 upsampling ratio. The final high-resolution belief maps $M^{HR}$ are then generated by the $\mathcal{PS}$ operation (see Fig. 4).

3.1 Belief Map Enhancement Network

Conventional pose estimation networks can not continuously in-
crease 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 $M^{LR}$ and $F^{LR}$, then conduct
a sequential regular convolution of $L - 1$ layers, and finally apply
an efficient sub-pixel convolution (the $L$th layer) that upscales the
low-resolution feature maps to high-resolution belief maps $M^{HR}$.

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

$$x = [F^{LR}, M^{LR}]$$  \hspace{1cm} (1)

$$f^L(x) = \text{ReLU}(w^L_T x)$$  \hspace{1cm} (2)

$$f^L(x) = \text{ReLU}(w^L_T f^{L-1}(x))$$  \hspace{1cm} (3)

Bases are absorbed in $w$ for simplicity. Here $[\cdot, \cdot]$ denotes concatenate.

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

$$M^{HR} = f^L(x) = \mathcal{PS}(w^L_T f^{L-1}(x))$$  \hspace{1cm} (4)

Where the weight $w_L$ 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

$$\mathcal{PS}(T)_{x,y:k} = T_{\lfloor x/r \rfloor, \lfloor y/r \rfloor, K \cdot r \mod(y, r) + K \cdot \mod(x, r) + k}$$  \hspace{1cm} (5)
Where \((x, y, k)\) represent coordinates in \(M^HR\) of size \(rH \times rW \times K\). Notice that the kernel size of \(w_k\) is \(3 \times 3\), which is greater than commonly used \(1 \times 1\) 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 \(z = \{z_k\}_{k=1}^K\), where \(z_k \in \mathbb{R}^2\) denotes the location of the \(k\)th keypoint of a person in the image. Then the high-resolution ground-truth belief map \(M^HR\) is generated from a Gaussian with mean \(z_k\) and standard deviation \(\sigma\):

\[
M^HR(p) \sim N(z_k, (\sigma^2)^2)
\]

(6)

Where \(p \in \mathbb{R}^2\) denotes the location, and \(\sigma\) 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 \(K\) belief maps, i.e., \(M^HR = \{M^HR_k\}_{k=1}^K\), for \(K\) body keypoints. We adopt Mean Squared Error (MSE) loss for model training. Given \(N\) input patches, the loss is defined by

\[
L^HR = \sum_{n=1}^{N} \sum_{k=1}^{K} ||M^HR - M^HR_k||^2
\]

(7)

Combined with the loss \(L^LR\) in base model, the total loss is

\[
L = L^LR + \eta L^HR
\]

(8)

Where \(\eta\) is the balance factor and we set \(\eta\) to 1 in all of our experiments.

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

We adopt sub-pixel convolution as the upsampling 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 used to increase resolution [10, 31, 34, 18]. However, deconvolution is also commonly used to increase resolution [10, 31, 34, 18].

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 upsampling factor \(r = 4\) in our EnhanceNet. For single person pose estimation the input patch size is \(256 \times 256\) and for multi person estimation the input patch size is \(256 \times 192\) except for Associative Embedding [26] whose patch size is \(512 \times 512\).

#### 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 Correct Keypoints) score as the evaluation metric. A keypoint is correct if it falls within \(\alpha\) pixels from the ground-truth location, where \(\alpha\) is the ground-truth head length and \(\alpha\) 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 \(\alpha\)).

<table>
<thead>
<tr>
<th>Methods</th>
<th>Head</th>
<th>Sho</th>
<th>Elb</th>
<th>Wri</th>
<th>Hip</th>
<th>Knee</th>
<th>Ank</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hourglass (2 stage) [27]</td>
<td>86.98</td>
<td>84.74</td>
<td>88.24</td>
<td>84.87</td>
<td>80.91</td>
<td>81.95</td>
<td>78.44</td>
<td>87.14</td>
</tr>
<tr>
<td>+EnhanceNet</td>
<td>93.70</td>
<td>95.14</td>
<td>89.13</td>
<td>84.00</td>
<td>87.35</td>
<td>84.12</td>
<td>79.38</td>
<td>87.96</td>
</tr>
<tr>
<td>Hourglass (4 stage)</td>
<td>96.49</td>
<td>95.50</td>
<td>89.89</td>
<td>84.46</td>
<td>87.43</td>
<td>84.65</td>
<td>80.21</td>
<td>88.34</td>
</tr>
<tr>
<td>+EnhanceNet</td>
<td>96.73</td>
<td>95.57</td>
<td>89.76</td>
<td>85.06</td>
<td>88.51</td>
<td>84.42</td>
<td>81.03</td>
<td>88.81</td>
</tr>
<tr>
<td>Hourglass (8 stage)</td>
<td>96.79</td>
<td>95.28</td>
<td>90.27</td>
<td>85.56</td>
<td>87.57</td>
<td>84.30</td>
<td>81.06</td>
<td>88.78</td>
</tr>
<tr>
<td>+EnhanceNet</td>
<td>96.79</td>
<td>95.41</td>
<td>90.30</td>
<td>85.41</td>
<td>88.14</td>
<td>84.85</td>
<td>81.25</td>
<td>89.03</td>
</tr>
<tr>
<td>DLCM [39]</td>
<td>96.75</td>
<td>95.01</td>
<td>90.33</td>
<td>86.76</td>
<td>89.14</td>
<td>86.90</td>
<td>83.55</td>
<td>90.37</td>
</tr>
<tr>
<td>+EnhanceNet</td>
<td>97.53</td>
<td>95.25</td>
<td>91.26</td>
<td>86.89</td>
<td>90.36</td>
<td>86.90</td>
<td>83.61</td>
<td>90.78</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Methods</th>
<th>Head</th>
<th>Sho</th>
<th>Elb</th>
<th>Wri</th>
<th>Hip</th>
<th>Knee</th>
<th>Ank</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>DLCM</td>
<td>96.8</td>
<td>95.2</td>
<td>89.3</td>
<td>84.4</td>
<td>88.4</td>
<td>83.4</td>
<td>78.0</td>
<td>88.5</td>
</tr>
<tr>
<td>+EnhanceNet</td>
<td>98.6</td>
<td>97.0</td>
<td>92.8</td>
<td>88.8</td>
<td>91.7</td>
<td>89.6</td>
<td>86.6</td>
<td>92.5</td>
</tr>
</tbody>
</table>

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 keypoint detection.

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.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Head</th>
<th>Sho</th>
<th>Elb</th>
<th>Wri</th>
<th>Hip</th>
<th>Knee</th>
<th>Ank</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>DLCM [39]</td>
<td>49.7</td>
<td>39.8</td>
<td>38.9</td>
<td>38.9</td>
<td>17.3</td>
<td>27.6</td>
<td>28.8</td>
<td>35.00</td>
</tr>
<tr>
<td>+EnhanceNet</td>
<td>53.84</td>
<td>42.48</td>
<td>43.43</td>
<td>40.67</td>
<td>18.40</td>
<td>29.34</td>
<td>31.44</td>
<td>37.55</td>
</tr>
</tbody>
</table>

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 points. We report standard average precisions and recall scores: AP, AP_{ok=0.50}, AP_{ok=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.

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. The APs are calculated on the COCO val2017 set.
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 $4\times 4$ zoom in).

(a) MPII

(b) MS COCO

**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.
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 key-point locations. Qualitative results on COCO are presented in Fig. 6b.

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. $M_{LR}$) 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 $F_{LR}$ can not bring more improvement than using $M_{LR}$, this indicates that our EnhanceNet mainly gains improvement by enhancing the belief maps. The best performance is achieved by concatenating $M_{LR}$ and $F_{LR}$, where the $F_{LR}$ 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 upsampling 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 upsampling 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 upsampling ratio of EnhanceNet since it has the best trade-off between model complexity and key-point 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


