

# Human Pose Estimation using Global and Local Normalization

Ke Sun<sup>1</sup>, Cuiling Lan<sup>2</sup>, Junliang Xing<sup>3</sup>, Wenjun Zeng<sup>2</sup>, Dong Liu<sup>1</sup>, Jingdong Wang<sup>2</sup>

<sup>1</sup> University of Science and Technology of China, Anhui, China   <sup>2</sup> Microsoft Research Asia, Beijing, China

<sup>3</sup> National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China

sunk@mail.ustc.edu.cn, {culan, wezeng, jingdw}@microsoft.com, jlxing@nlpr.ia.ac.cn, dongeliu@ustc.edu.cn

## Abstract

In this paper, we address the problem of estimating the positions of human joints, i.e., articulated pose estimation. Recent state-of-the-art solutions model two key issues, joint detection and spatial configuration refinement, together using convolutional neural networks. Our work mainly focuses on spatial configuration refinement by reducing variations of human poses statistically, which is motivated by the observation that the scattered distribution of the relative locations of joints (e.g., the left wrist is distributed nearly uniformly in a circular area around the left shoulder) makes the learning of convolutional spatial models hard. We present a two-stage normalization scheme, human body normalization and limb normalization, to make the distribution of the relative joint locations compact, resulting in easier learning of convolutional spatial models and more accurate pose estimation. In addition, our empirical results show that incorporating multi-scale supervision and multi-scale fusion into the joint detection network is beneficial. Experiment results demonstrate that our method consistently outperforms state-of-the-art methods on the benchmarks.

## 1. Introduction

Human pose estimation is one of the most challenging problems in computer vision and plays an essential role in human body modeling. It has wide applications such as human action recognition [35], activity analyses [1], and human-computer interaction [29]. Despite many years of research with significant progress made recently [3, 11, 10, 8, 32, 31], pose estimation still remains a very challenging task, mainly due to the large variations in body postures, shapes, complex inter-dependency of parts, clothing and so on.

This work was done when Ke Sun was an intern at Microsoft Research Asia. Junliang Xing is partly supported by the Natural Science Foundation of China (Grant No. 61672519).

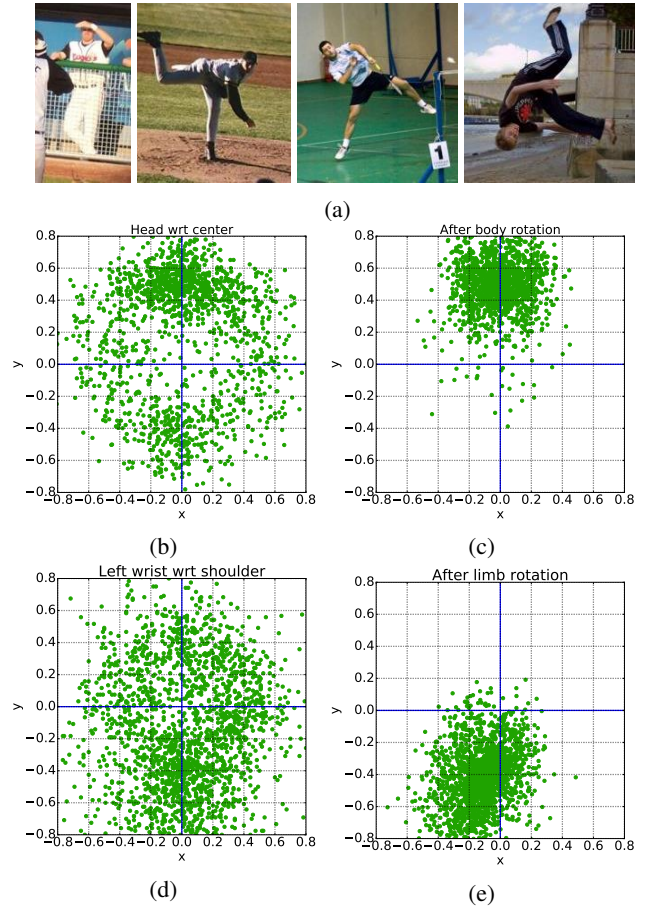


Figure 1: Pose normalization can compact the relative position distribution. (a) Example images with various poses from LSPET [17]. (b) and (d) are originally relative positions of two joints. (b) shows the positions of heads with respect to the body centers. (d) shows the positions of the left wrists with respect to the left shoulders. (c) and (e) are the relative positions after body and limb normalization corresponding to (b) and (d) respectively. The distributions of the relative positions in (c) and (e) are much more compact.

There are two key problems in pose estimation: joint detection, which estimates the confidence level of each pixel being some joint, and spatial refinement, which usually refines the confidence for joints by exploiting the spatial configuration of the human body.

Our work follows this path and mainly focuses on spatial configuration refinement. It is observed that the distribution of relative positions of joints may be very diverse with respect to their neighboring ones. Examples regarding the distributions of the joints on the LSPET dataset [17] are shown in Figure 1. The relative positions of the head with respect to the center of the human body are shown in Figure 1 (b), which is distributed almost uniformly in a circular region. After making the human body upright, the distribution becomes much more compact, as shown in Figure 1 (c). We have similar observations for other neighboring joints. For some joints on the limbs (*e.g.*, wrist, ankle), their distributions are still diverse even after positioning the torso upright. We further rotate the human upper limb (*e.g.*, the left arm) to a vertical downward positions. The distribution of the relative positions of the left wrist, shown in Figure 1 (e), becomes much more compact.

The diversity of orientations (*e.g.*, body and limb) is the main factor in the variations of pose. Motivated by these observations, we propose two normalization schemes, reducing diversity to generate compact distributions. The first normalization scheme is human body normalization, rotating the human body to upright according to joint detection results, which globally makes the relative positions between joints compactly distributed. This scheme is followed by a global spatial refinement module to refine all the estimations of the joints. The second one is limb normalization: rotating the joints of each limb to make the relative positions more compact. There are four total limb normalization modules, and each is followed by a spatial limb refinement module to refine the estimations from the global spatial refinement. Thanks to the normalization schemes, a much more consistent spatial configuration of the human body can be obtained, which facilitates the learning of spatial refinement models.

Besides the observations in [20, 21, 30, 36] that the multi-stage supervision, *e.g.*, supervision on the joint detection stage and the spatial refinement stage, is helpful, we observe that multi-scale supervision and multi-scale fusion over the convolutional network within the joint detection stage are also beneficial.

Our main contribution lies in effective normalization schemes to facilitate the learning of convolutional spatial models. Our scheme can be applied following different joint detectors for refining the spatial configurations. Our experiment results demonstrate the effectiveness on several joint detectors, such as FCN [21], ResNet [13] and Hourglass [22]. An additional minor contribution is that we em-

pirically show the improvement by using an architecture with multi-scale supervision and fusion for joint detection.

## 2. Related Work

Significant progress has been made recently in human pose estimation by deep learning based methods [33, 23, 8, 31, 25, 14, 22]. The joint detection and joint relation models are widely recognized as two key components in solving this problem. In the following, we briefly review related developments on these two components respectively and discuss some related works which motivate our design of the normalization scheme.

**Joint detection model.** Many recent works use convolutional neural networks to learn feature representations for obtaining the score maps of joints or the locations of joints [33, 12, 6, 34, 22, 26, 5]. Some methods directly employ learned feature representations to regress joint positions, *e.g.*, the DeepPose method [33]. A more typical way of joint detection is to estimate a score map for each joint based on the fully convolutional neural network (FCN) [21]. The estimation procedure can be formulated as a multi-class classification problem [34, 22] or regression problem [6, 31]. For the multi-class formulation, either a single-label based loss (*e.g.*, softmax cross-entropy loss) [9] or a multi-label based loss (*e.g.*, sigmoid cross-entropy loss) [25] can be used. One main problem for the FCN-based joint detection model is that the positions of joints are estimated from low resolution score maps. This reduces the location accuracy of the joints. In our work, we introduce multi-scale supervision and fusion to further improve performance with gradual up-sampling.

**Joint relation model.** The pictorial structures [38, 24] define the deformable configurations by spring-like connections between pairs of parts to model complex joint relations. Subsequent works [23, 8, 37] extend such an idea to convolutional neuron networks. In those approaches, to model the human poses with large variations, a mixture model is usually learned for each joint. Tompson et al. [32] formulates the spatial relations as a Markov Random Field (MRF) like model over the distribution of spatial locations for each body part. The location distribution of one joint relative to another is modeled by convolutional prior which is expected to give some spatial predictions and remove false positive outliers for each joint. Similarly, the structured feature learning method in [9] adapts geometrical transform kernels to capture the spatial relationships of joints from feature maps. To better estimate the human pose, complex network architecture design with many more parameters are expected on account of the articulated structure of the human body, such as [9] and [37]. In our work, we address this problem by compensating for the variations of poses both globally and locally for facilitating the spatial configuration exploration.

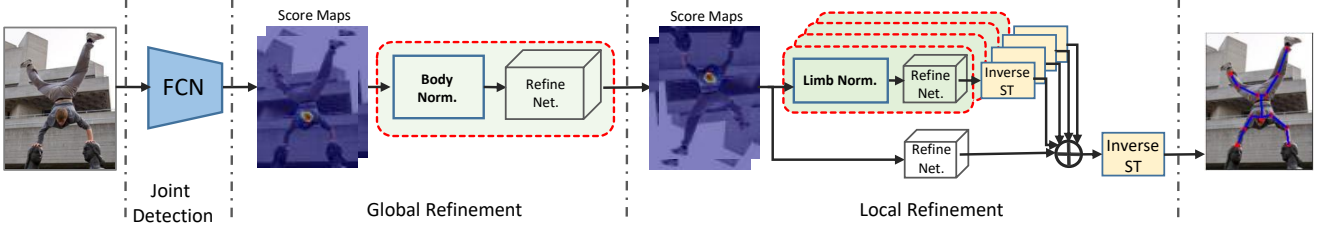


Figure 2: Proposed framework with global and local normalization. Joint detection with fully convolutional network (FCN) provides initial estimation of joints in terms of score maps. In the global refinement stage, a body (global) normalization module rotates the score maps to have upright position for the body, followed by a refinement module. In the local refinement stage, limb (local) normalization modules rotates the score maps to have vertical downward position for limbs, followed by refinements.

**Normalization.** Normalization of the training samples to reduce their variations has been proven to be a key step before learning models using these samples. For example, in the PCA Whitening and ZCA Whitening operations [18], the feature pre-processing step are adapted before training an SVM classifier, etc. The batch normalization [15] technique accelerates the deep network training by reducing the internal covariate shift across layers. In computer vision applications, face normalization has been found to be very helpful for improving the face recognition performance [4, 41, 7]. It is beneficial for decreasing the intra-person variations and achieving pose-invariant recognition.

### 3. Our Approach

Human pose estimation is defined as the problem of localization of human joints, *i.e.*, head, neck *et al.*, from an image. Given an image  $I$ , the goal is to estimate the positions of the joints:  $\{(x_k, y_k)\}_{k=1}^K$ , where  $K$  is the number of joints.

#### 3.1. Pipeline

Figure 2 shows the framework of our proposed approach. It consists of joint detection and spatial configuration refinement, which are both realized with convolutional neural networks. The output of joint detector consists of  $K + 1$  score maps, including  $K$  joint score maps, providing spatial configuration information and one non-joint (background) score map. The value in each score map indicates the degree of confidence that the pixel is the corresponding joint. With the score maps generated by the former stage (*e.g.*, joint detector or refinement stage) as the input, two normalization stages correct wrongly predicted joints based on spatial configurations of the human body. Note that we focus on the exploration of the spatial configurations of joints for refinement. Unlike many other works [5, 22, 34], we do not incorporate low level features and our refinement is based on the score maps which indicate the probabilities of being each joint.

#### 3.2. Spatial Configuration Refinement

There are two stages for spatial configuration refinement as depicted in Figure 2. The first stage is a global refinement, consisting of a global normalization module and a refinement module that refines all  $K$  joints. The second stage includes two parallel refinement modules: semi-global refinement and local refinement. The local refinement module consists of four branches. Each branch corresponds to a limb and contains a local limb normalization module and a local refinement module. Inverse normalizations by inverse spatial transforms are used to rotate the joints/body back for obtaining the final results.

**Body normalization.** The purpose of body normalization is to make the orientation of the whole body the same, *e.g.*, upright in our implementation<sup>1</sup>. Specifically, we rotate the body as well as the  $K$  score maps around the center of the four joints (*i.e.*, left shoulder, right shoulder, left hip, right hip) so that the line from the center to the neck joint is upright, as shown in Figure 3 (b). The positions of the joints are estimated from the  $K$  Gaussian-smoothed score maps by finding the maximum responses in each map and returning the corresponding position as the position of the joint.

We implement the normalization through spatial transform, which is written as follows,

$$\bar{\mathbf{x}} = \mathbf{R}(\mathbf{x} - \mathbf{c}) + \mathbf{c}, \quad (1)$$

where  $\mathbf{c}$  is defined as the center of the four joints on the torso,  $\mathbf{c} = \frac{1}{4}(\mathbf{p}_{l-shoulder} + \mathbf{p}_{r-shoulder} + \mathbf{p}_{l-hip} + \mathbf{p}_{r-hip})$ ,  $\mathbf{p}_{l-shoulder}$  denotes the estimated location of the left shoulder joint, and  $\mathbf{R}$  is a rotation matrix,

$$\mathbf{R} = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}. \quad (2)$$

Here  $\theta = \arccos \frac{(\mathbf{p}_{neck} - \mathbf{c}) \cdot \mathbf{e}_{\perp}}{\|\mathbf{p}_{neck} - \mathbf{c}\|_2}$ ,  $\mathbf{e}_{\perp}$  denotes the unit vector along the vertical upward direction, which is illustrated in Figure 3.

<sup>1</sup>Essentially, any orientation is fine in our approach.

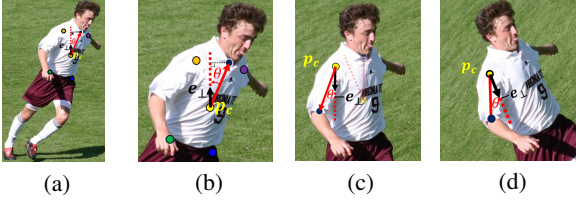


Figure 3: Illustration of the body normalization and limb normalization. (a) and (b) show the rotation angle  $\theta$  for body normalization. (c) is the image after body normalization and the rotation angle for a limb normalization. (d) shows the image after limb normalization. Note that our network actually performs the normalization on the score maps rather than on the image. From (b) to (d), we show the magnified view of the images for clarity.

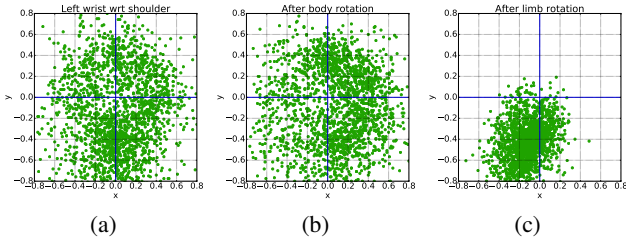


Figure 4: Limb normalization can compact the relative position distribution for some joints on limbs, which is hard to address via body normalization. (a), (b) and (c) are the relative positions of left wrist with respect to the position of left shoulder, and the relative positions after body normalization, and that after limb normalization. The distribution in (c) is much more compact.

**Local normalization.** The end joints on the four limbs have higher variations. As illustrated in Figure 4 (a) and (b), through body normalization, the distribution of the wrist with respect to the shoulder is still not compact. Limb normalization is then adopted where we rotate the arm to have upper arm vertical downwards, with the distribution, as shown in Figure 4 (c), becoming much more compact. There are four local normalization modules corresponding to the four limbs respectively. Each limb contains three joints: a root joint (shoulder, hip), a middle joint (elbow, knee), and an end joint (wrist, ankle). We perform the normalization by rotating the corresponding three score maps around the root joint such that the line connecting the root joint and the middle joint has a consistent orientation, *e.g.*, vertical downwards in our implementation. The normalization process is illustrated in Figures 3 (c) and (d).

**Discussions.** There are some alternative solutions for handling the diverse distribution problem of relative locations of the joints, *e.g.*, type supervision [9], and mixture model [39]. To check the effectiveness of our proposed nor-

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total	#param.
FCN	93.3	86.7	74.4	68.0	85.7	82.0	78.5	81.5	134M
Type-supervision	93.5	87.5	76.4	68.8	87.8	82.8	79.5	82.3	(134+11)M
Multi-branch	<b>93.8</b>	87.7	76.3	69.4	87.7	82.6	79.8	82.5	(134+33)M
Global normalization	93.7	<b>88.8</b>	<b>77.3</b>	<b>69.6</b>	<b>88.4</b>	<b>84.0</b>	<b>81.0</b>	<b>83.3</b>	(134+11)M

Table 1: Comparing our global normalization-based solution with type supervision and the multi-branch solution on LSP dataset with the OC annotation (@PCK 0.2) trained on the LSP dataset.

malization scheme, we provide two alternative solutions by considering the diversity of pose types. Here, we obtain pose type information by clustering human poses into three types from the LSP dataset.

In our first alternative solution (*type-supervision*), based on our global refinement framework, we remove the normalization model but add type supervision in the refinement network by learning three sets of score maps (*i.e.*  $3 \times K+1$ ) rather than one set. In the second alternative solution (*multi-branch*), based on our global refinement framework, we remove the normalization model but extend the refinement network to multi-branches, with each branch handling one type of pose. Note that the number of parameters for the three-branch spatial configuration refinement is three times of ours. Specifically, for the alternative two solutions, we process the training data with extra data augmentation, to make the number of training data for each type similar to ours. The multi-branch approach is computationally more expensive and requires more training time than ours.

We take the original FCN [21] as the joint detector, and make a comparison among the two alternative solutions and our global normalization refinement scheme. Table 1 shows the results. The two alternative solutions improve performance over FCN, but under-performs our approach. It is possible to further improve the performance of the multi-branch approach with more extensive data augmentation and more branches, but this will increase the computational complexity for both training and testing. In addition, our approach can also benefit from the multi-branch and type supervision solutions, where our normalization is applied to each branch and the type supervision further constrains the degrees of freedom of parts.

Figure 5 shows examples of estimated poses from different stages of our network (Figure 5 (a), (c), (d)), and that from the similar network but without normalization modules (*i.e.*, Figure 5 (b)). With global and local normalization, joint estimation accuracy is much improved (*e.g.*, knee on the top images, ankle on the bottom images). Our normalization scheme can reduce the diversity of human poses and facilitate the inference process. For example, the consistent orientation of the left hip and the left knee makes it easier to infer the location of the left ankle on the bottom example in Figure 5.



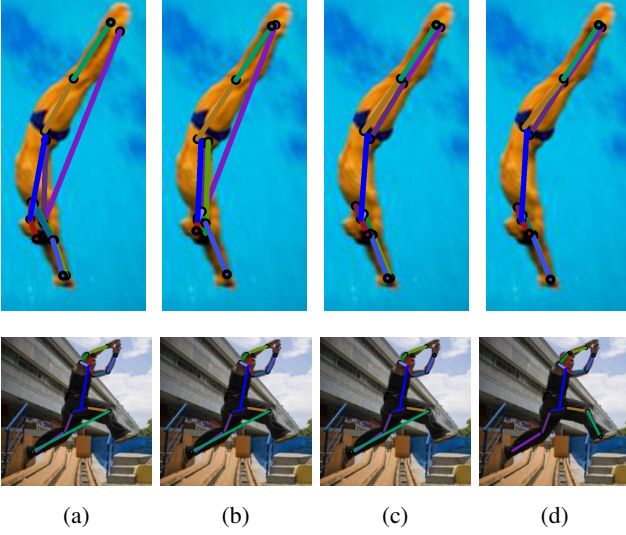


Figure 5: Estimated poses from (a) FCN, (b) the scheme with spatial refinement but without body normalization and limb normalization, (c) our scheme with body normalization, (d) our scheme with both body normalization and limb normalization.

### 3.3. Multi-Scale Supervision and Fusion for Joint Detection

To efficiently train the FCN and exploit intermediate-level representations, we introduce multi-scale supervision and multi-scale fusion, which show performance gain in many works [20, 21, 30, 36]. The network structure is provided in Figure 6. Multi-scale supervision makes the network concentrate on accurate localization on different resolutions, avoiding loss in accuracy due to down-sampling. This is different from [22, 34], adding multi-supervision to each stage with the same resolution. Multi-scale fusion exploits the information at different scales. More details are introduced in the Section 3.4.

### 3.4. Implementation Details

**Network architectures.** The proposed network architecture contains three main parts: the base network for joint detection, the normalization network, and the refinement network.

**Joint detection network:** We use fully convolutional network as our joint detector. For fairness of comparison, we use architectures similar to the compared methods as the joint detectors. We demonstrate the effectiveness of our normalization scheme on top of different joint detectors: the improved FCN as showed in Figure 6, ResNet-152 similar to [14, 5] and Hourglass [22]. The FCN generates three sets of score maps (FCN\_32s, FCN\_16s, and FCN\_8s), at different resolutions, corresponding to the last three deconvolution layers with strides 32, 16, and 8 respectively. The

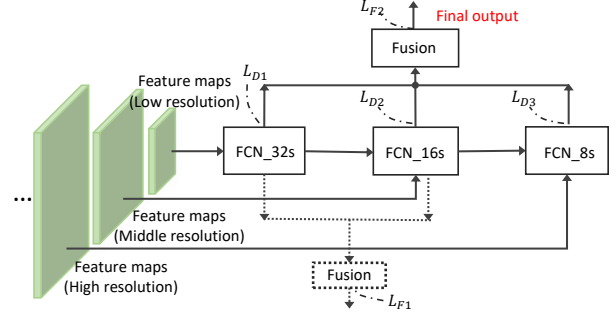


Figure 6: Architecture of the improved FCN. We utilize multi-scale supervision and fusion.

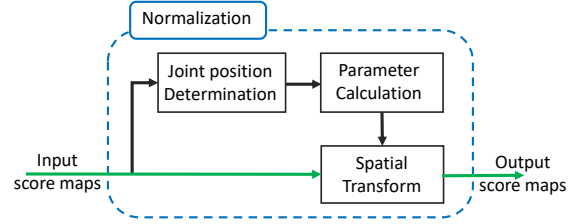


Figure 7: Normalization module. Based on the score maps, a module derives the joint positions. Then the spatial transform parameters can be calculated directly.

fusion is an ensemble of different scales with a  $1 \times 1$  convolutional layer. We introduce multi-scale supervision (with losses  $L_{D1}$ ,  $L_{D2}$ , and  $L_{D3}$ ) and multi-scale fusion (with losses  $L_{F1}$ ,  $L_{F2}$ ) (see Figure 6). The architecture of Hourglass is the same as [22] and ResNet is similar to [14].

**Normalization network:** In Figure 7, we show the flowchart of the normalization module. The spatial transform is performed on the score maps with the calculated transform parameters. End-to-end training is supported with the error back propagation along the transform path (as denoted by the green line). For the joint position determination module, a Gaussian blur is performed on the mapped score maps, with the mapping corresponding to the Sigmoid like operation or no operation, depending the loss design of the joint detection network. Then, the position corresponding to the maximal value in each processed score map is estimated as the position of that joint. The network calculates the rotation center  $c$  and the rotation angle  $\theta$  based on the estimated positions of joints. All the operations are incorporated into the network as layers.

**Refinement network:** The refinement network consists of four convolutional layers. The convolutional kernel sizes and channel numbers for the four layers are  $9 \times 9 \times (K+1)$  with 128 output channels,  $15 \times 15 \times 128$  with 128 output channels,  $15 \times 15 \times 128$  with 128 output channels, and  $1 \times 1 \times 128$  with  $J$  output channels, where  $J$  denotes the number of output joints. Large kernel sizes of  $9 \times 9$  and  $15 \times 15$  are beneficial for capturing spatial information of each joint.

**Loss functions.** The groundtruth is generated to be  $K$  score maps. When FCN or ResNet detector is utilized, the pixels in a circled region with radius of  $r$  centered at a joint are labeled by 1 while other pixels are set by 0. We define the radius  $r$  as 0.15 times of the distance between left shoulder and right hip. When Hourglass detector is utilized, 2D Gaussian centered on the joint location is used for groundtruth labeling [22]. For the spatial refinement stages, both visible and occluded joints are labeled.

For the FCN joint detector, softmax function is utilized to estimate the probability of being some joint for the visible joints. For spatial refinement, after the several convolution layers, sigmoid-like operation  $1/(1 + e^{-(wx+b)})$  is used to map the score  $x$  to the estimation of the probability of being some joint (both visible and occluded joints). Here,  $w$  and  $b$  are two learnable parameters which transform the scores to be in a suitable range of the domain of sigmoid function. For other joint detectors, *i.e.* ResNet and Hourglass, we use their designed loss.

**Optimization.** We pretrain the network by optimizing joint detector, global and local refinement model, and then fine-tune the whole framework.

We initialize the parameters of the refinement model randomly with a Gaussian distributed variable of variance 0.001. When the network of joint detector converges, we fix it and train the body refinement network with a base learning rate of 0.001. Afterwards, we fix the former networks and train the limb refinement network. Finally, we fine-tune the entire network with learning rate 0.0002.

For FCN, we initialize it with the model weights from PASCAL VOC [21]. During training, we progressively minimize the loss function: first minimize  $L_{D1}$  then  $L_{D1} + L_{D2} + L_{F1}$ , then  $L_{D1} + L_{D2} + L_{D3} + L_{F1} + L_{F2}$  (see Figure 6). FCN detector is implemented based on Caffe and SGD is taken as the optimization algorithm. The initial learning rate is set to 0.001. For other joint detectors, such as ResNet-152 and Hourglass, we adopt the same settings as proposed by the authors in their papers.

## 4. Experiments

**Datasets.** We evaluate the proposed method on four datasets: Leeds Sports Pose (LSP) [16], extended LSP (LSPET) [17], Frames Labeled in Cinema (FLIC) [28] and MPII Human Pose [2]. The LSP dataset contains 1000 training and 1000 testing images from sports activities, with 14 full body joints annotated. The LSPET dataset adds 10,000 more training samples to the LSP dataset. The FLIC dataset contains 3987 training and 1016 testing images with 10 upper body joints annotated. The MPII Human Pose dataset includes about 25k images with 40k annotated poses. Existing works evaluate the performance on the LSP dataset with different training data, which we follow for per-

formance comparisons respectively.

**Evaluation criteria.** The metrics “Percentage of Correct Keypoints (PCK)” and the “Area Under Curve (AUC)” are utilized for evaluation [39, 25]. A joint is correct if it falls within  $\alpha \cdot l_r$  pixels of the groundtruth position, with  $\alpha$  denoting a threshold and  $l_r$  a reference length.  $l_r$  is the torso length for the LSP, FLIC, and the head size for the MPII.

**Data Augmentation.** For the LSP dataset, we augment the training data by performing random scaling with a scaling factor between 0.80 and 1.25, horizontal flipping, and rotating the data across 360 degrees, in consideration of its unbalanced distribution of pose orientations. All input images are resized to  $340 \times 340$  pixels. For the FLIC and the MPII dataset, we randomly rotate the data across  $\pm 30$  degrees and resize images into  $256 \times 256$  pixels.

### 4.1. Results

We denote our final model as *Ours(detector+Refine)*, and the scheme with only detector as *Ours(detector)*. All the experiments are conducted without any post-processing.

**LSP OC.** With the LSP dataset as training data, Table 2 shows the comparisons with the OC annotation for per-joint PCK results, the overall results at threshold  $\alpha = 0.2$  (@PCK0.2), and the AUC. Our method achieves the best performance, where the AUC is 4.3% higher than Chu *et al.* [9], even though the layer number of their additional network (180 conv layers) is much larger than our refinement network (24 conv layers). Our refinement provides 1% improvement in the overall accuracy.

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total	AUC
Kiefel et al. [19]	83.5	73.7	55.9	36.2	73.7	70.5	66.9	65.8	38.6
Ramakrishna et al. [27]	84.9	77.8	61.4	47.2	73.6	69.1	68.8	69.0	35.2
Pishchulin et al. [24]	87.5	77.6	61.4	47.6	79.0	75.2	68.4	71.0	45.0
Ouyang et al. [23]	86.5	78.2	61.7	49.3	76.9	70.0	67.6	70.0	43.1
Chen&Yuille [8]	91.5	84.7	70.3	63.2	82.7	78.1	72.0	77.5	44.8
Yang et al. [37]	90.6	<b>89.1</b>	<b>80.3</b>	73.5	85.5	82.8	68.8	81.5	43.4
Chu et al. [9]	93.7	87.2	78.2	<b>73.8</b>	88.2	83.0	<b>80.9</b>	83.6	50.3
Ours(FCN)	94.3	87.8	77.1	69.8	87.1	83.7	79.7	82.8	54.2
Ours(FCN+Refine)	<b>94.9</b>	88.8	77.6	70.7	<b>88.9</b>	<b>84.8</b>	80.5	<b>83.7</b>	<b>54.6</b>

Table 2: Performance comparison on the LSP testing set with the OC annotation (@PCK0.2) trained on the LSP training set.

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total	AUC
Pishchulin et al. [25]	<b>97.4</b>	<b>92.0</b>	83.8	<b>79.0</b>	93.1	88.3	83.7	88.2	<b>65.0</b>
Ours(FCN)	96.2	90.7	83.3	77.5	91.2	89.3	85.0	87.6	61.8
Ours(FCN+Refine)	96.7	91.8	<b>84.4</b>	78.3	<b>93.3</b>	<b>90.7</b>	<b>85.8</b>	<b>88.7</b>	63.0

Table 3: Performance comparison on the LSP testing set with the OC annotation (@PCK0.2) trained on the MPII+LSPET+LSP training set.

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total	AUC
Tompson et al. [32]	90.6	79.2	67.9	63.4	69.5	71.0	64.2	72.3	47.3
Fan et al. [12]	92.4	75.2	65.3	64.0	75.7	68.3	70.4	73.0	43.2
Carreira et al. [6]	90.5	81.8	65.8	59.8	81.6	70.6	62.0	73.1	41.5
Chen&Yuille [8]	91.8	78.2	<b>71.8</b>	65.5	73.3	70.2	63.4	73.4	40.1
Yang et al. [37]	90.6	78.1	73.8	<b>68.8</b>	74.8	69.9	58.9	73.6	39.3
Ours(FCN)	93.8	80.3	69.7	64.7	81.0	78.1	73.1	77.2	50.5
Ours(FCN+Refine)	<b>94.0</b>	<b>80.9</b>	70.6	65.3	<b>82.3</b>	<b>78.5</b>	<b>73.7</b>	<b>77.9</b>	<b>50.7</b>

Table 4: Performance comparison on the LSP testing set with PC (@PCK0.2) trained on the LSP training set.

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total	AUC
Bulat et al. [5]	<b>98.4</b>	86.6	79.5	73.5	88.1	83.2	78.5	83.5	—
Wei et al. [34]	—	—	—	—	—	—	—	84.32	—
Rafi et al. [26]	95.8	86.2	79.3	75.0	86.6	83.8	79.8	83.8	56.9
Yu et al. [40]	87.2	88.2	<b>82.4</b>	<b>76.3</b>	<b>91.4</b>	<b>85.8</b>	78.7	84.3	55.2
Ours(FCN)	95.2	86.2	78.1	72.8	87.0	85.7	81.3	83.7	56.1
Ours(FCN+Refine)	95.5	<b>88.5</b>	80.0	73.9	89.8	<b>85.8</b>	<b>81.5</b>	<b>85.0</b>	<b>58.5</b>

Table 5: Performance comparison on the LSP testing set with PC (@PCK0.2) trained on the LSP+LSPET training set.

Another work [25] incorporates the MPII and LSPET dataset for training. The results with the same training set are shown in Table 3. Our refinement achieves 1.1% improvement in overall accuracy and outperforms the start-of-the-art even though our detector does not use location refinement and an auxiliary task as used by [25].

**LSP PC.** Table 4 shows the comparisons with PC annotation. Compared with the result of Yang *et al.* [37] on the LSP dataset, our method significantly improves the performance by 4.3% in overall accuracy and 11.4% in AUC.

We incorporate the LSPET dataset into the training data and evaluate the performance with PC annotation. From Table 5, we can see that our scheme achieves the best performance. Yu *et al.* [40] extracts many pose bases to represent various human poses. In contrast, our method normalizes various poses. Our method outperforms theirs by 3.3% in AUC and 0.7% in the overall accuracy.

To verify the effectiveness of our normalization scheme, we connect our refinement model at the end of those deeper joint detectors, *i.e.*, ResNet-152 (152 layers) [14, 5] and Hourglass (about 300 layers) [22]. The results are shown in Table 6. Without using the location refinement and auxiliary task [14], our baseline scheme *Ours(ResNet-152)* drops about 1% than [14]. With the proposed refinement added, our scheme *Ours(ResNet+Refine)* improves over the baseline by 1% in the overall accuracy and is comparable to [14]. Bulat *et al.* [5] added a modified hourglass network [22] (90 layers, with parameters being three times larger than our refinement model) after ResNet-152. *Ours(ResNet+Refine)*

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total	AUC
Insafutdinov et al. [14]	97.4	92.7	87.5	84.4	91.5	89.9	87.2	90.1	<b>66.1</b>
Wei et al. [34]	97.8	92.5	87.0	83.9	91.5	90.8	89.9	90.5	65.4
Bulat et al. [5]	97.2	92.1	88.1	85.2	92.2	<b>91.4</b>	88.7	90.7	63.4
Ours(ResNet-152)	97.0	91.5	86.2	82.8	89.4	89.9	87.5	89.2	63.5
Ours(ResNet+Refine)	97.3	92.2	87.1	83.5	92.1	90.6	87.8	90.1	64.8
Ours(Hourglass)	97.7	93.0	88.3	84.8	92.3	90.2	90.0	90.9	65
Ours(Hg+Refine)	<b>97.9</b>	<b>93.6</b>	<b>89.0</b>	<b>85.8</b>	<b>92.9</b>	91.2	<b>90.5</b>	<b>91.6</b>	65.9

Table 6: Performance comparison on the LSP testing set with the PC annotation (@PCK0.2) trained on the MPII+LSPET+LSP training set.

	Head	Shoulder	Elbow	Wrist	AUC
Toshev et al. [33]	—	—	92.3	82	—
Tompson et al. [32]	—	—	93.1	89	—
Chen&Yuille.[8]	—	—	95.3	92.4	—
Wei et al. [34]	—	—	97.6	95	—
Newell et al. [22]	—	—	99.0	97.0	—
ResNet-152	99.7	99.7	99.1	97	75.3
Ours(Refine)	<b>99.9</b>	<b>99.8</b>	<b>99.5</b>	<b>97.7</b>	<b>76.9</b>

Table 7: Performance comparison on the FLIC dataset with OC annotation (@PCK0.2).

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Hourglass [22]	<b>98.2</b>	<b>96.3</b>	<b>91.2</b>	87.1	<b>90.1</b>	<b>87.4</b>	83.6	90.9
Ours(Hourglass+Refine)	98.1	96.2	<b>91.2</b>	<b>87.2</b>	89.8	<b>87.4</b>	<b>84.1</b>	<b>91.0</b>

Table 8: Performance comparison on the MPII test set (@PCKh0.5) trained on the MPII training set.

is 1.4% better than that of Bulat *et al.* [5] in AUC. When we take Hourglass as our detector, the proposed refinement brings 0.5% improvement in the overall accuracy.

**FLIC dataset.** We evaluate our method on the FLIC dataset with the OC annotation. We take ResNet-152 [13] as our joint detection network. Table 7 shows that our refinement improves over the baseline model by 0.4% for elbow, 0.7% for wrist, and 1.6% in AUC.

**MPII dataset.** We take Hourglass [22] as our joint detector and evaluate our method on the MPII dataset. Table 8 shows that our refinement performs similarly on the test set in overall accuracy. On the validation set, we obtains 0.4% improvement. To check the reason for small gains, we analyze the relative position distribution on the MPII validation dataset. We found that the original distribution without normalization is already compact, being similar to the distribution after the normalization on the LSP dataset. Unlike the poses in the LSP dataset (sport poses), the majority of poses are upright and normal, as shown in Figure 8. Our normalization scheme presents its advantages on the datasets including high diverse poses. In reality, these complicated postures are inevitable.

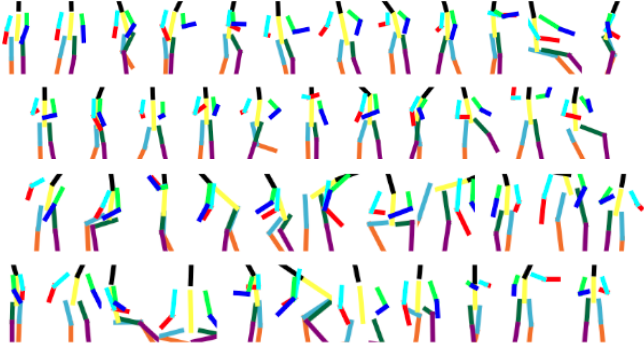


Figure 8: Body pose clusters on the MPII test set. The maker of MPII dataset clusters body poses into 45 types on the test set. Note the figure is from <http://human-pose.mpi-inf.mpg.de/#results>.

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
FCN	<b>94.3</b>	87.8	77.1	69.8	87.1	83.7	79.7	82.8
Stage-1 w/o body norm.	93.5	88.2	77.2	69.8	87.5	83.8	80.2	82.9
Stage-1 w body norm.	94.2	<b>88.5</b>	<b>77.8</b>	69.8	88.2	83.8	80.2	83.2
Stage-2 w/o limb norm.	94	88.3	77.7	69.4	88	83.8	80.3	83
Stage-2 w limb norm.	94.2	88.4	77.7	<b>70.4</b>	<b>88.8</b>	<b>84.7</b>	<b>80.5</b>	<b>83.5</b>

Table 9: Evaluation of body normalization and limb normalization on the LSP test dataset with the OC annotation (@PCK0.2) trained on the LSP training dataset.

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
FCN	94.9	90.5	82.2	74.8	89	88.2	83.5	86.2
Stage-1 w/o body norm.	94.8	90.1	81.6	75.2	90	87.7	83	86
Stage-1 w body norm.	94.9	90.8	83.8	76.3	89.7	88.3	84	86.8
Stage-2 w/o limb norm.	95	88.3	80.6	75.7	88.5	86.2	82.8	85.3
Stage-2 w limb norm.	<b>95.4</b>	<b>91.1</b>	<b>84</b>	<b>76.8</b>	<b>90.9</b>	<b>89</b>	<b>84.4</b>	<b>87.4</b>

Table 10: Evaluation of body normalization and limb normalization on the LSP test dataset with the OC annotation (@PCK0.2) trained on the LSP+LSPET training dataset.

## 4.2. Ablation Study

We analyze the effectiveness of the proposed components, including the two pose normalization and refinement stages, and the multi-scale supervision and fusion.

**Global and local normalization.** To verify the effectiveness of body and limb normalization, we compare the results of the network with normalization versus that without normalization on the two stages separately.

Table 9 shows the comparisons on the LSP dataset. With body normalization, the shoulder is 0.7% higher than that of FCN and the hip estimation is improved by 1.1%. In contrast, the model without the body normalization introduces much smaller improvement. With limb normalization, the accuracy of wrist, knee, and ankle is improved by 0.6%,

	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total	AUC
FCN_32s	93.7	85.2	74.4	65.2	86.2	81.2	77	80.4	50
FCN_16s	93.9	85.9	75.1	68.3	86.3	83.4	78.5	81.6	52.1
FCN_8s	94.2	86.2	75.8	68.8	86.5	83.8	78.3	82	53.7
FCN_16s (Extra)	94.2	87.5	76.8	69.2	87.5	82.8	78.4	82.4	53.2
FCN_16s (Fusion)	<b>94.3</b>	87.7	77.0	69.5	87.6	83.4	78.6	82.6	53.2
FCN_8s (Extra)	94.2	87.5	<b>77.2</b>	69.6	<b>87.2</b>	83.5	<b>79.7</b>	82.6	<b>54.2</b>
FCN_8s (Fusion)	<b>94.3</b>	<b>87.8</b>	77.1	<b>69.8</b>	<b>87.7</b>	<b>83.7</b>	<b>79.7</b>	<b>82.8</b>	<b>54.2</b>

Table 11: Evaluation of multi-scale supervision and multi-scale fusion on top of FCN on the LSP testing set with the OC annotation (@PCK0.2) trained on the LSP training set.

0.9%, and 0.3% respectively. Without pose normalization, the subnetwork tends to preserve the results of the former stage. Similar phenomena are observed when we use the LSP+LSPET dataset for training as shown in Table 10. We notice that the performance of Stage-1 without body normalization even provides interior performance than FCN. In contrast, when body normalization is utilized, consistent performance improvement can be achieved.

**Multi-scale supervision and fusion.** For FCN, we add multi-scale supervision and multi-scale score map fusion to improve accuracy. Here, we evaluate the efficiency of the extra supervision and fusion respectively. Table 11 shows the experiment results. FCN\_16s and FCN\_8s denote the results of the original FCN without extra loss and fusion at the middle and high resolution respectively. FCN\_16s (Extra) and FCN\_8s (Extra) denote the results after adding supervision. FCN\_16s (Fusion) and FCN\_8s (Fusion) denote the results after adding both supervision and fusion. From Table 11, we have the following two observations. First, with extra supervision, the accuracy of most joints improves by more than 1% and the AUC increases noticeably at the same resolution level. Note that FCN\_8s (Extra) achieves similar accuracy as FCN\_16s (Extra) but its AUC is much higher. Second, we fuse the score maps together with different weights to exploit their respective advantages. We can see the overall accuracy improves by a further 0.2%.

## 5. Conclusion

In this paper, considering that the distributions of the relative locations of joints are very diverse, we propose a two-stage normalization scheme: human body normalization and limb normalization, making the distributions compact and facilitating the learning of spatial refinement models. To validate the effectiveness of our method, we connect the refinement model to various state-of-the-art joint detectors. Experiment results demonstrate that our method consistently improves the performance on different benchmarks.



## References

- [1] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *ACM Computing Surveys*, 43(3):16, 2011.
- [2] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014.
- [3] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, 2009.
- [4] A. Athana, T. K. Marks, M. J. Jones, K. H. Tieu, and M. Rohith. Fully automatic pose-invariant face recognition via 3D pose normalization. In *ICCV*, 2011.
- [5] A. Bulat and G. Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *ECCV*, 2016.
- [6] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human pose estimation with iterative error feedback. In *CVPR*, 2016.
- [7] D. Chen, G. Hua, F. Wen, and J. Sun. Supervised transformer network for efficient face detection. In *ECCV*, 2016.
- [8] X. Chen and A. Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *NIPS*, 2014.
- [9] X. Chu, W. Ouyang, H. Li, and X. Wang. Structured feature learning for pose estimation. In *CVPR*, 2016.
- [10] M. Dantone, J. Gall, C. Leistner, and L. Van Gool. Human pose estimation using body parts dependent joint regressors. In *ICCV*, 2013.
- [11] M. Eichner, V. Ferrari, and S. Zurich. Better appearance models for pictorial structures. In *BMVC*, volume 2, page 5, 2009.
- [12] X. Fan, K. Zheng, Y. Lin, and S. Wang. Combining local appearance and holistic view: Dual-source deep neural networks for human pose estimation. In *CVPR*, 2015.
- [13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.
- [14] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepcut: A deeper, stronger, and faster multi-person pose estimation model. In *ECCV*, 2016.
- [15] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pages 448–456, 2015.
- [16] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010.
- [17] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR*, 2011.
- [18] A. Kessy, A. Lewin, and K. Strimmer. Optimal whitening and decorrelation. *arXiv preprint arXiv:1512.00809*, 2015.
- [19] M. Kiefel and P. V. Gehler. Human pose estimation with fields of parts. In *ECCV*, 2014.
- [20] C.-Y. Lee, S. Xie, P. W. Gallagher, Z. Zhang, and Z. Tu. Deeply-supervised nets. In *AISTATS*, 2015.
- [21] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.
- [22] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016.
- [23] W. Ouyang, X. Chu, and X. Wang. Multi-source deep learning for human pose estimation. In *CVPR*, 2014.
- [24] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Poselet conditioned pictorial structures. In *CVPR*, 2013.
- [25] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *CVPR*, 2016.
- [26] U. Rafi, J. Gall, and B. Leibe. An efficient convolutional network for human pose estimation. In *BMVC*, 2016.
- [27] V. Ramakrishna, D. Munoz, M. Hebert, A. Bagnell, and Y. Sheikh. Pose machines: Articulated pose estimation via inference machines. In *ECCV*, 2014.
- [28] B. Sapp and B. Taskar. Modoc: Multimodal decomposable models for human pose estimation. In *CVPR*, 2013.
- [29] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 56(1):116–124, 2013.
- [30] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.
- [31] J. J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler. Efficient object localization using convolutional networks. In *CVPR*, 2015.
- [32] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *NIPS*, 2014.
- [33] A. Toshev and C. Szegedy. DeepPose: Human pose estimation via deep neural networks. In *CVPR*, 2014.
- [34] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, 2016.
- [35] B. Xiaohan Nie, C. Xiong, and S.-C. Zhu. Joint action recognition and pose estimation from video. In *CVPR*, 2015.
- [36] S. Xie and Z. Tu. Holistically-nested edge detection. In *ICCV*, 2015.
- [37] W. Yang, W. Ouyang, H. Li, and X. Wang. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation. In *CVPR*, 2016.
- [38] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011.
- [39] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE Trans. Pattern Anal. and Mach. Intell.*, 35(12):2878–2890, 2013.
- [40] X. Yu, F. Zhou, and M. Chandraker. Deep deformation network for object landmark localization. In *ECCV*, 2016.
- [41] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li. High-fidelity pose and expression normalization for face recognition in the wild. In *CVPR*, 2015.