Hierarchical Contextual Refinement Networks for Human Pose Estimation

Xuecheng Nie, Jiashi Feng, Junliang Xing, Shengtao Xiao, Shuicheng Yan, *Fellow, IEEE*

***Abstract*—Predicting human pose in the wild is a challenging problem due to high flexibility of joints and possible occlusion. Existing approaches generally tackle the difficulties either by holistic prediction or multi-stage processing, which suffer from poor performance for locating challenging joints or high computational cost. In this paper, we propose a new Hierarchical Contextual Refinement Network (HCRN) to robustly predict human poses in an efficient manner, where human body joints of different complexities are processed at different layers in a context hierarchy. Different from existing approaches, our proposed model predicts positions of joints from easy to difficult in a single stage through effectively exploiting informative contexts provided in the previous layer. Such approach offers two appealing advantages over state-of-the-arts: (1) more accurate than predicting all the joints together and (2) more efficient than multi-stage processing methods. We design a Contextual Refinement Unit (CRU) to implement the proposed model, which enables auto-diffusion of joint detection results to effectively transfer informative context from easy joints to difficult ones. In this way, difficult joints can be reliably detected even in presence of occlusion or severe distracting factors. Multiple CRUs are organized into a tree-structured hierarchy which is end-to-end trainable and does not require processing joints for multiple iterations. Comprehensive experiments evaluate the efficacy and efficiency of the proposed HCRN model to improve well-established baselines and achieve new state-of-the-art on multiple human pose estimation benchmarks.**

***Index Terms*—Human Pose Estimation, Joint Complexity-Aware, Hierarchical Contextual Refinement Network.**

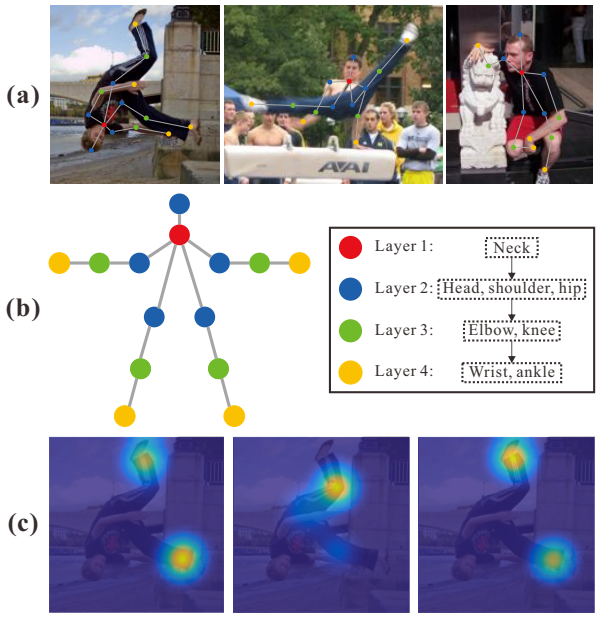


Fig. 1. (a) Example images from LSP dataset showing various complexities of body joints. (b) The proposed complexity-aware hierarchy of body joints. Body joints are divided into four layers according to their complexities. (c) Left: initial heatmap of left ankle. Middle: heatmap of left knee providing clues for localizing left ankle. Right: refined heatmap of left ankle. The proposed approach can use contextual information from easier joints to guide the estimation of more difficult joints in a single stage.

I. Introduction

HUMAN pose estimation aims to generate joint configurations of human body from a single image. It is a fundamental task in computer vision, with wide application in AR/VR [1], [2], gaming [3], human computer interaction [4], and human behavior analysis [5], [6], [7]. In literature [8], [9], [10], [11], [12], efforts have been made to tackle various challenges in human pose estimation, *e.g.*, high degree of freedom articulations, view-point changes, occlusion, self-similarities and large pose and appearance variations.

Despite significant progress made by those approaches, some challenges have not been well addressed. One of them is from the heterogeneous flexibilities and complexities of human body joints. For example, as shown in Fig. 1 (a), neck and head top of human body are relatively easier to estimate than other parts due to their lower degrees of freedom and strong discriminative characteristics. Comparatively, wrist and ankle are much more difficult because of higher degrees of freedom, occlusion, and ambiguities. Most existing approaches, however, ignore the difference of such joint complexities and stick to dealing with all joints together in a holistic way, thus the difficult joints may contaminate and then degrade their performance. Based on this observation, to further improve the performance of human pose estimation, we propose to explicitly take the complexities of joints into consideration to avoid the above problem encountered by holistic approaches.

In particular, we propose to divide the huge space of possible joint configurations of human body into smaller ones according to the joint complexities. With this divide-and-conquer strategy, estimation can be performed for joints within each smaller space alternatively. With such motivation, we devise a complexity-aware hierarchical model that distributes body joints into different layers according to their complexities. In the proposed hierarchy, easy joints are estimated first and difficult ones are addressed later utilizing the estimation results for easier ones. Specifically, all body joints are divided into 4 layers as shown in Fig. 1 (b). Layer 1, acting as the root

X. Nie, J. Feng, and S. Yan are with the Learning and Vision Lab, ECE Department, National University of Singapore, Singapore 117583. S. Yan is also with Qihoo 360 AI Institute, China. E-mail: niexuecheng@u.nus.edu, elefjia@nus.edu.sg and yanshuicheng@360.cn

J. Xing is with National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, China. E-mail: jlxing@nlpr.ia.ac.cn

S. Xiao is with Artificial Intelligence Institute at Qihoo 360 International. E-mail: xstgavin124@gmail.com

of the proposed hierarchical model, only includes neck which is generally believed to be the easiest joint to localize. Layer 2 includes 5 joints, *i.e.*, head, left/right shoulder and left/right hip, which are of slightly greater flexibility. Left/right elbow and left/right knee are put at layer 3. Layer 4 contains left/right wrist and left/right ankle, which are the most difficult joints for estimation. The intuition behind such a division of body joints is based on the degree of freedom from kinematics of human pose [13] and appearance discrimination. This way can help mitigate negative effects of difficult joints upon the estimation of easy ones. It can also alleviate estimation difficulties in the smaller joint configuration space. Moreover, our complexity-aware hierarchical model can be adapted to different body joint annotations, *e.g.,* it can utilize the skeleton definition proposed by [14] if one dataset provides the torso center annotation.

In addition to the complexity dependence within a layer of the proposed hierarchical model, articulated body joints in neighboring layers are also closely correlated spatially. Our key observation is that easy joints can provide useful contextual information for localizing difficult ones. As shown in Fig. 1 (c), the left image shows a wrongly estimated heatmap for left ankle due to its ambiguity with right ankle; the middle image is the correctly estimated heatmap for left knee. Because of the close spatial correlations, left knee can provide informative guidance to refine the heatmap of left ankle into a correct one as shown in right image in Fig. 1 (c). The refinement process can be conducted recurrently from easy joints to difficult ones. Whereas, current approaches often adopt stacked or iterative multi-stage networks for heatmap reuse to implicitly explore contextual information, *e.g.* Convolutional Pose Machine [12], Iterative Error Feedback [15] and Chained Prediction Model [16], resulting in low efficiency, massive auxiliary parameters or suffering from error accumulation.

Motivated by this, we propose a new Contextual Refinement Unit (CRU) to effectively aggregate contextual information provided by easy joints and help estimate difficult ones and generate more accurate heatmap for each body joint. We organize multiple CRUs into a tree-structure to capture various joint complexities and form the Hierarchical Contextual Refinement Network (HCRN) for efficient human pose estimation layer-by-layer. In stark contrast with the iterative refinement or chain models [12], [15], [16], HCRN can accurately predict joint locations within only one stage, benefiting from the layered structure and the proposed CRU.

We use a front-end Convolutional Neural Network (CNN) to learn deep representations, which is integrated with the proposed HCRN model into a unified framework for end-to-end training and inference. Extensive experiments on benchmarks LSP [17], FLIC [18], MPII Human Pose Single-Person [19] and MSCOCO [20] have shown superior performance and efficiency of the proposed HCRN model. Our contributions can be summarized into three aspects: 1) We propose a principled way to deal with heterogeneous complexities of human body joints. 2) We introduce a new neural network unit, called the CRU, which can effectively integrate and exploit spatial contextual information from well-estimated joints. 3) We propose a Hierarchical Contextual Refinement Network that consists of stacked CRUs with a tree-structure. We apply it to address human pose estimation problem, achieving superior performance and high efficiency over well-established baselines and new state-of-the-art.

II. Related Work

In literature, human pose estimation in monocular images has been heavily relied on pictorial structure model. [21] and [22] have adopted this model via constructing an undirected graph where nodes represents joints and edges their relationships. Later, [23] introduced poselets to encode high-order joint dependencies. [24] exploited deformable part models for articulated human detection and pose estimation. [25] and [26] utilized loopy constraints to address the double-counting problem encountered with tree-structure models. However, they are severely limited with hand-craft features, such as SIFT, HOG, etc. Recently, Convolutional Neural Networks (CNNs) have been used to replace hand-crafted features with deep learned ones, significantly pushing the frontier of human pose estimation and becoming the dominant stream.

Existing CNN based approaches can be roughly divided into two categories: CNN with classification model and CNN with regression model. CNN with classification model approaches predict the confidence maps, which indicates probabilities of body joints presenting at each position in the images. Follow this strategy, [27], [11], [28], [29] have integrated CNNs with graphical models, *e. g.*, [11] has proposed a spatial-model to mimic Markov Random Field (MRF) loopy belief propagation. [30] has presented a two-stage cascaded CNN architecture, which first outputs part detection heatmaps to provide part attention and contextual information and then performs part regression to predict the part location in the image. [31] has presented a CNN based recurrent model for iteratively increasing the receptive field of the network via combining the intermediate feature representations to learn contextual information and improve the final heatmap predictions. In [32], Chu *et al.* have proposed to integrate CNN with a multi-context attention mechanism for human pose estimation, through generating heatmaps from features at multiple resolutions with various semantics to keep global consistency of full human body and local accuracy of single body joints. In [33], Chen *et al.* has exploited the adversarial training strategy via introducing a discriminator to implicitly constrain poses from the generator.

The CNN with regression model approaches solve human pose estimation by directly regressing the joint positions or the offsets between the initial positions of joints and the ground truth positions. To solve the highly non-linear mapping problem involved in the regression, CNN with regression model approaches often use an iterative updating mechanism to progressively reduce the regression error and get closer to ground truth. For such iterative regression approaches, at each iteration they estimate the corrections that should be made on current predictions to reduce the difference from the ground truth. Then the correction parameters are added to current predictions to refine the estimation results. For instance, Liu *et al.* [34] have presented a three-stage cascaded CNN framework for human pose estimation via predicting the coarse-level

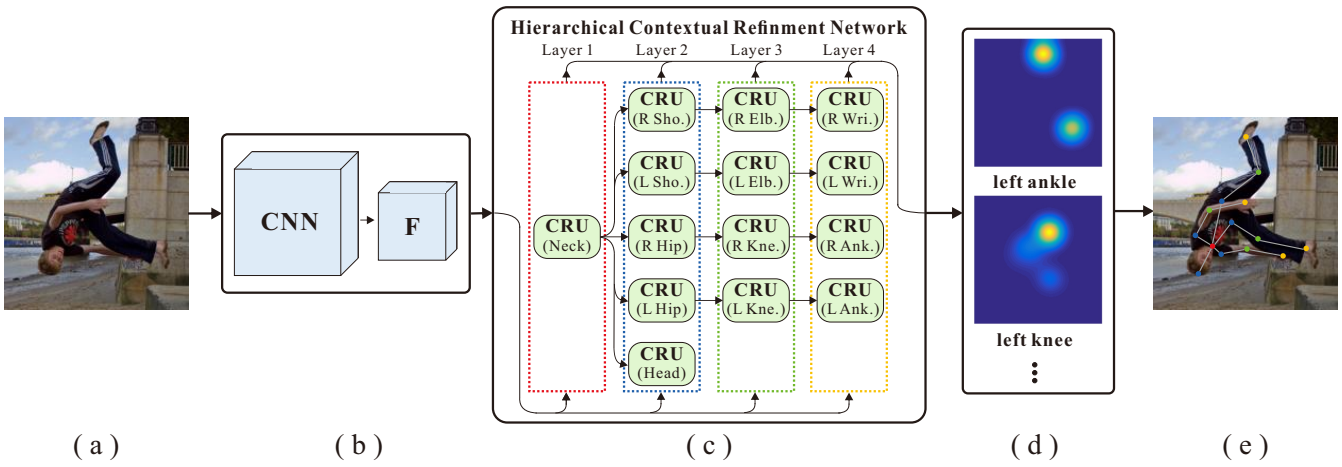


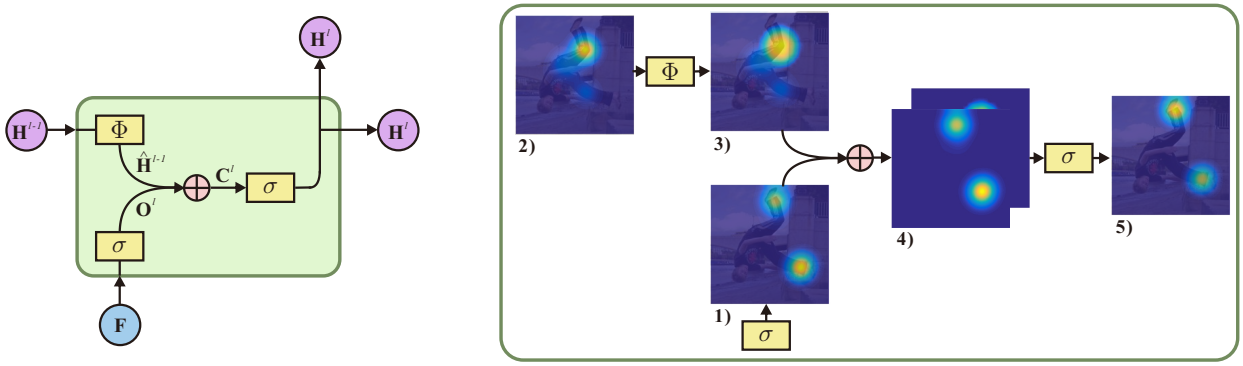
Fig. 2. Overview of the proposed approach. (a) The input image. (b) A front-end CNN for learning representations of body joints. (c) The proposed Hierarchical Contextual Refinement Network for estimating human pose sequentially layer-by-layer according to the complexities of joints, from easy to difficult. (d) The generated joint-specific heatmaps. (e) The final estimation results.

joint locations in the first stage and then estimating joint offsets to groundtruth in the following two stages to refine the joint locations. Recurrent Neural Networks (RNNs) [35], [9], especially Long-Short Term Memory (LSTM) [36], [37], [38], have been used to directly model the connections between each iteration. In [39], Belagiannis *et al.* have proposed a robust optimization scheme to utilize the Tukeys weight function as loss instead of L2 loss for CNN models to improve regression accuracy via considering outliers in training samples.

Different from those approaches which holistically classify or regress all joints together, in this work we propose to estimate human body joints layer-by-layer sequentially according to their complexities, and exploit results from an easy layer for getting effective guidance for estimation of joints in a more difficult layer. This approach can simplify the estimation process and improve the performance substantially.

Some existing works also consider structuring human body parts. However, the existing hierarchical models for human pose estimation construct part correlation structures according to some simple clues, *e.g.*, scales and sizes. For instance, Sun *et al.* [40] have represented human body as a collection of parts from coarse-to-fine, where fine-level child joints are spatially connected to their coarse-level parent joints, such as the relationships between upper/lower arm and entire arm. Then, they formulate the parent-children relationships recursively to model the consistency of whole body configurations. Similarly, Tian *et al.* [41] have proposed a hierarchical model by dividing human body into five parts consisting of primitive joints. They built pairwise relations between nearby parts and joints, resulting in a tree-structure model. Then the objective function is formulated as MRF model and solved by the message passing algorithm. In [42], Ionescu *et al.* have proposed to generate joint descriptors for 3D human pose estimation based on 2D body part annotations via performing second-order label-sensitive pooling over region hierarchies, where the coarsest level contains a region for the entire body and the finest level has different regions for each body part. Although this kind of hierarchical model can simplify joint detection based on guidance from large parts, it does not build correlations among joints directly or deal with complexities of different joints explicitly. Fan *et al.* [14] have proposed the pose locality constrained representation to improve 3D human pose estimation from monocular images via hierarchically dividing human pose space into low-dimension spaces using subspace clustering to explicitly encourage pose locality in human-body modeling, whereas, ignoring complexity differences of joints. Gkioxari *et al.* [16] have proposed a chained prediction model to sequentially estimate joints according to their complexities. However, their approach requires to exploit CNNs multiple times for estimating all joints, leading to accumulation of prediction error during the iterative propagation process.

In hand pose estimation, hierarchical models formulated according to joint complexities have been widely used. Sun *et al.* [43] have proposed a hierarchical model to estimate hand pose sequentially. They first estimate the location of palm, the easiest joint to predict, then fix the location of palm and used the palm information to guide estimation of more complex fingers. Tang *et al.* [44] have proposed to regress hand pose progressively in four layers from wrist, metacarpophalangeal, proximal interphalangeal, to distal interphalangeal and tips, according to the complexities of different joints. Estimation results from one layer can help simplify estimation of joints in the next layer. Ye *et al.* [45] also have proposed a hierarchical approach for hand pose estimation combining discriminative approach and generative approach, where the whole hand is divided into four layers from palm to tips. With the spatial attention mechanism, they divide both input and output spaces into hierarchies. In each layer, cascaded CNNs are used to estimate the joints progressively. Results from prior layers reduce the search space of next layers. Comprehensive experiments conducted by these methods have shown superior performance of a hierarchical model than the traditional holistic estimation. Motivated by previous works, we propose a new complexity-aware hierarchical model for human pose estimation. It is formulated according to the complexity of a single joint, and can explicitly model the relationships between joints, rather than compositions of joints utilized in [40] and [41] that implicitly model such relationships.



(a) (b)

Fig. 3. Illustration of the proposed Contextual Refinement Unit (CRU). (a) The structure of CRU. (b) An example for CRU: 1) the initial heatmap of left ankle, note that the heatmap is wrongly responded on right ankle; 2) the estimated heatmap of left knee for the given image; 3) the diffused heatmap of left knee; 4) the concatenated heatmaps; 5) the refined heatmap of left ankle.

III. Complexity-Aware HCRNs

A. Model Overview

We propose a Hierarchical Contextual Refinement Network (HCRN) to localize joints in a complexity-aware way. The overall architecture of the proposed HCRN for human pose estimation is shown in Fig. 2. Our goal is to estimate the positions $\mathcal{P} = \{P_i\}_{i=1}^N$ of N joints $\mathcal{J} = \{J_i\}_{i=1}^N$ for the given human body image I, *e.g.*, Fig. 2 (a). We use P_i to denote the 2D location of joint J_i, *i.e.*, $P_i = (x_i, y_i)$. For joint localization, we first utilize a front-end CNN to learn discriminative representation from the image, which encodes the high-level semantic information enabling the following classifiers to predict probabilities on the presence of body joints in each location, denoted as F shown in Fig. 2 (b). Then F is fed into the HCRN to generate heatmaps $\mathcal{H} = \{H_i\}_{i=1}^N$ for all joints. Each element in one heatmap H_i indicates the possibility of the corresponding location containing the i-th joint. The proposed HCRN model estimates location heatmaps of different joints sequentially from low layers to high ones according to its internal complexity-aware hierarchy. The information from easy joints will be utilized for localizing more difficult ones in the following layer. To fully integrate the information across layers, we introduce a new Contextual Refinement Unit (CRU). Intuitively, CRUs refine the initial heatmaps of difficult joints estimated from F by exploiting the guidance of contextual heatmaps generated by diffusing probability in each location of input heatmaps. As shown in Fig. 2 (c), CRUs are organized into a complexity-aware hierarchy of human body joints, forming a hierarchical contextual refinement network. In this way, HCRN applies CRUs layer-by-layer recursively to generate heatmaps for all joints as shown in Fig. 2 (d). The position P_i for joint J_i is localized by taking the location with the maximum confidence score on heatmap H_i. Fig. 2 (e) shows the final estimation result for the given input image. Details of the proposed CRU will be given in the next subsection.

B. Contextual Refinement Units

The structure of CRU is shown in Fig. 3 (a). It estimates heatmap H_i^l for joint J_i^l in the layer l according to the deep representation F and the heatmap H_{i*}^{l-1} of the corresponding joint J_{i*}^{l-1} in the layer $(l-1)$. For simplicity, we hide the subscripts in the following specifications. CRU conducts four steps of calculations to update the estimations on joint heatmaps which can be expressed as

$$O^l = \sigma(W_O^l * F + B_O^l), \tag{1}$$

$$\hat{H}^{l-1} = \Phi(H^{l-1}), \tag{2}$$

$$C^l = O^l \oplus \hat{H}^{l-1}, \tag{3}$$

$$H^l = \sigma(W_C^l * C^l + B_C^l), \tag{4}$$

where $*$ denotes the convolution operator, $\oplus$ denotes the concatenation operator, $\sigma(\cdot)$ is the sigmoid activation function, and $\Phi(\cdot)$ is a probability diffusion function, which will be detailed below. The CRU is parameterized by W_O^l, B_O^l, W_C^l and B_C^l that are end-to-end learnable.

Step 1: Heatmap Initialization The first step of CRU defined in Eqn. (1) is to estimate initial heatmap O^l for a joint J^l in layer l. In this step, a 1×1 convolutional operation defined by parameters $W_O^l \in \mathbb{R}^{c\times1\times1}$ and $B_O^l \in \mathbb{R}^{1\times1\times1}$ is performed on the input feature map F to generate the dense responses for joint J^l, and c is the channel dimension of F. Then, the sigmoid operation is performed on the generated response map to give the initial heatmap O^l. An example of initial heatmap O^l for left ankle in layer 4 is shown in Fig. 3 (b) 1), from which one can see that the initial heatmap cannot provide reliable estimation for left ankle due to the false alarm on right ankle caused by ambiguities of these two joints of symmetry.

Step 2: Heatmap Diffusion The second step of CRU defined in Eqn. (2) aims to spread the confidence score of each position in the heatmap H^{l-1} from the previous layer to its neighbors, such that CRU can aggregate contextual information and guide the estimation of H^l. In this step, CRU performs probability diffusion operation $\Phi(\cdot)$ based on deformation information of H^{l-1} to generate the diffused heatmap $\hat{H}^{l-1}$. For a pixel P in H^{l-1}, the probability diffusion operator $\Phi(\cdot)$ is defined as

$$\Phi(H^{l-1}(P)) = \max_{\delta\in\mathcal{N}} \left(H^{l-1}(P+\delta) - W_d^{l-1} d(\delta)\right), \tag{5}$$

where $\delta = (\delta x, \delta y)$ is the position offset, $\mathcal{N} = [-r, r]\times[-r, r]$ is the range of δ defining the fusion field ($r = 7$ in our

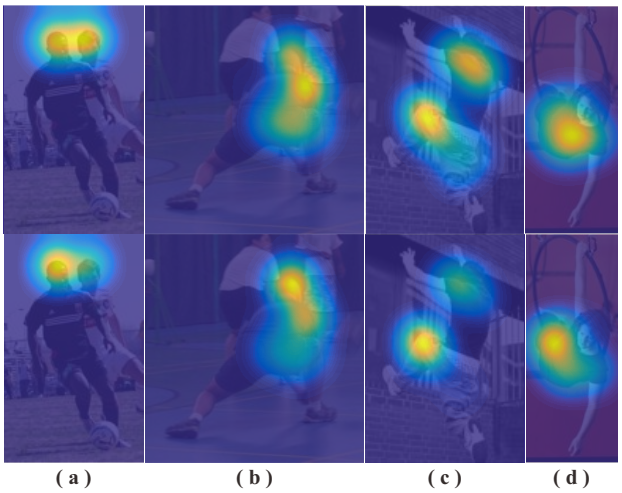


Fig. 4. Examples of the proposed CRU dealing with various false alarm situations occurred in initial heatmaps. The first row shows the initial heatmaps and the second row shows the refined heatmaps. (a) shows the refinement for heatmap of head through dealing with false alarm caused by overlapping people. (b) and (c) show refinements for heatmaps of head and left knee through dealing with false alarms caused by inaccurate estimations, respectively. (d) shows the refinement for heatmap of left knee through dealing with false alarm caused by occlusion.

experiments), $d(\delta) = [\delta x, \delta y, \delta x^2, \delta y^2]$ is the deformation feature, and W_d^{l-1} is a 4D weight vector. By definition of Eqn. (5), the score of each location in heatmap H^{l-1} is spread to its neighbors according to the deformation features, which will provide powerful contextual information for the prediction of H^l. This operation is implemented as a probability diffusion layer in Caffe [46] framework for end-to-end training and testing. An example of probability diffusion operation is shown in Fig. 3 (b). Given the heatmap H^{l-1} of left knee in Fig. 3 (b) 2), after probability diffusion, the diffused heatmap $\hat{H}^{l-1}$ is shown in Fig. 3 (b) 3).

Step 3: Heatmap Stacking The third step of CRU defined in Eqn. (3) is to stack the initial heatmap O^l with the transformed heatmap $\hat{H}^{l-1}$, which can be used in the fourth step of CRU to add the contextual information from $\hat{H}^{l-1}$ to O^{l-1} for estimating H^l. This step is implemented by a Concat layer. An example of stacked heatmaps is given in Fig. 3 (b) 4).

Step 4: Heatmap Refinement The fourth step of CRU defined in Eqn. (4) is to refine the initial heatmap O^l by selecting proper contextual information from $\hat{H}^{l-1}$ to generate the estimation result H^l for joint J^l in the layer l. This operation is implemented by a convolutional operation on the stacked heatmaps C^l followed by a sigmoid operation. Parameters $W_C^l \in \mathbb{R}^{2\times1\times1}$ and $B_C^l \in \mathbb{R}^{1\times1\times1}$ of convolution can be automatically learned during the training process. An example of the final estimated heatmap H^l for left ankle is shown in Fig. 3 (b) 5). We can see false alarm for left ankle is reduced and heatmap is corrected by guidance from left knee.

From the above four steps, CRU can refine the initial estimation O^l of a difficult joint relying on the contextual information from an easy joint H^{l-1}. More examples for the refinement process of CRU are shown in Fig. 4. From the figure, we can see that the CRU can deal with many false alarm cases that occur in initial heatmaps caused by overlapping people, inaccuracy, and occlusion.

C. The Front-End CNN

To learn discriminative representation in the image, we adopt four different kinds of networks in our experiments: VGG network with 16 layers (VGG16) [47], Residue network with 101 layers (Res101) [48], Convolution Pose Machines with 6 stages (CPM) [12] and Hourglass network with 8 stages (HG) [10], motivated by the recent progress in deep learning. We extract the feature $F \in \mathbb{R}^{c\times h\times w}$ from "fc7" of VGG16, "res5c" of Res101, and highest-level features of the last stage of CPM as well as HG, as input to HCRN, respectively, where c is the dimension of discriminative representation and h and w are the spatial-size height and width of F. Previous works have shown that VGG16 and Res101 with stride of 32 pixels are too coarse to precisely locate joints. Hence, we adopt the dilation algorithm [49] to reduce the stride to 8 pixels for improving localization accuracy. We initialize VGG16 and Res101 with ImageNet pretrain models. For CPM and HG, we follow the same settings as the original papers [12], [10]. We use Sigmoid activation function on predictions and cross entropy loss. We add supervision in both the steps of Heatmap Initialization and Heatmap Refinement of CRU. We train our model using Caffe [46] with SGD as the optimization algorithm for VGG16, ResNet101 and CPM based models, Pytorch with RMSProp [50] for HG based ones.

IV. Experiments

We conduct experiments to evaluate the proposed model on three single-person pose estimation benchmarks LSP [17], FLIC [18] and MPII Human Pose Single-Person [19], and one multi-person pose estimation benchmark MSCOCO [20]. Details are illustrated in the following subsections.

A. Experiments on LSP Dataset

We first evaluate our model on the Leeds Sports Pose (LSP) dataset [17]. It contains 2,000 images about sport activities with challenging articulations, including 1,000 images for training and 1,000 images for testing. Each person in the LSP dataset is roughly 150 pixels in height with full-body annotation of 14 joints. In addition, we also use the Leeds Sports Pose Extended Training (LSPET) dataset [51] in our training, which includes 10,000 images collected from Flicrk.com with the same configurations as the LSP dataset. Therefore, we get 11,000 images for training our model and baselines.

We further augment training data with rotation degrees in $[-40°, 40°]$, scaling with factors in [0.7, 1.3] and horizontally mirror. These training samples are resized and padded to 368×368 pixels. The target heatmap for each joint is constructed by assigning a positive label 1 at each location within 10 pixels to the ground truth, and otherwise a negative label 0. We also construct a target heatmap for background to get additional supervision for model training. We train the network on LSP dataset for 250 epochs in total. We test our model on 3-scale image pyramids to deal with scale variations and compare with baselines on the LSP testing set. We use two widely used metrics, Percentage of Correct Keypoints (PCK) [52] and Percentage of Correct Parts (PCP) [27] for evaluation on LSP with both Observer-Centric (OC) [53]

TABLE I
EXPERIMENTS ON LSP DATASET WITH PCK PC METRIC. THE PERFORMANCE OF OUR PROPOSED MODELS (VGG16-, RES101-, CPM- AND HG-HCRN), TWO ABLATED VERSIONS (VGG16- AND RES101-HCRN-W/O-DIFFUSION), TWO BASELINES (VGG16- AND RES101-HOLISTIC), AND STATE-OF-THE-ART MODELS CPM AND HG, ARE PRESENTED FOR COMPARISON.

Methods	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Avg	Time in secs.
HG-HCRN	**97.6**	**88.0**	**82.0**	**76.1**	**88.5**	**85.4**	**82.6**	**85.7**	0.202
HG	97.5	87.7	81.7	75.5	88.2	85.0	82.0	85.4	0.196
Res101-HCRN	96.8	86.8	79.7	75.8	85.7	84.8	80.1	84.2	0.188
Res101-HCRN-w/o-Diffusion	96.4	86.2	79.0	73.7	85.6	83.3	78.6	83.3	0.184
Res101-Holistic	96.3	85.8	78.2	72.1	86.4	82.5	77.5	82.7	0.182
VGG16-HCRN	95.9	85.5	78.0	72.1	82.9	81.3	73.6	81.3	0.067
VGG16-HCRN-w/o-Diffusion	95.5	85.3	77.2	69.7	82.0	80.4	71.1	80.2	0.063
VGG16-Holistic	95.4	85.1	76.5	68.7	81.4	79.7	69.9	79.5	0.061
CPM-HCRN	97.0	87.1	81.1	75.9	87.2	84.5	82.0	85.0	0.283
CPM-6-Stage	96.9	86.7	80.4	74.7	86.7	83.3	80.3	84.1	0.277
CPM-3-Stage	96.7	84.8	78.6	73.5	85.8	82.1	74.8	82.3	0.232
CPM-2-Stage	96.3	83.3	75.8	69.6	82.2	76.0	62.9	78.0	0.220
CPM-1-Stage	93.6	69.7	54.8	53.8	56.3	55.6	52.9	62.4	0.103

TABLE II
COMPARISON OF THE PARAMETER NUMBER OF HCRN AND ONE STAGE OF CPM.

Methods	ParamNum
HCRN	98
One Stage CPM	6,563,712

and Person-Centric (PC) [54] settings. With PCP, a predicted body part is considered correct if its segment endpoints lie within 50% of the length of the ground-truth segment from their annotated location. PCK defines a predicted keypoint to be correct if it falls within $\alpha \cdot \max(H, W)$ pixels of the groundtruth keypoint, where H and W are the height and width of the torso bounding box, respectively, and α controls the relative threshold for considering correctness, empirically set as α=0.2. PC and OC indicate two different annotation settings for defining right and left body parts of persons in images. PC means that right/left body parts are marked according to the viewpoint of the person in the image, while OC the viewpoint of the observer.

Ablation Analysis We first perform ablation analysis of the proposed model and results are shown in Table I. We use VGG16-, Res101-, CPM- and HG-HCRN to denote the proposed hierarchical contextual refinement networks with different backbones, respectively. We use VGG16-Holistic and Res101-Holistic to denote the holistic estimation directly from the front-end CNN. We also compare our model with CPM and HG, the state-of-the-art models on LSP dataset. We reimplement CPM and HG based on the codes provided by the authors, and the performance is very close to the reported ones in their papers. We compare the performance of different models under the PCK PC metric. We also list speed for one inference of these models on a single TITAN X GPU.

From Table I, we can see that with the multi-stage design, CPM continuously improves its performance when increasing the number of stages, from 62.4% in the first stage to 84.1% in the sixth stage. However, the time cost for inference is almost tripled from 0.103 seconds to 0.277 seconds. The increment of time cost for CPM to add one more layer is about 0.013 seconds averagely. In contrast to CPM, the single stage VGG16-HCRN network achieves much higher accuracy of 81.3% compared with CPM-1-Stage, with only half of the time cost. The high efficiency and accuracy of VGG16-HCRN are from (1) the HCRN component effectively refines prediction of difficult joints and (2) more efficient network design that uses smaller kernel size and deeper CNN model. In contrast, CPM network uses larger kernels in order for getting sufficiently large receptive fields, which slows down the inference process. The heatmap diffusion across different joints within HCRN effectively avoids this inefficient design by introducing the hierarchical heatmap diffusion. In addition, we also compare the parameter number of HCRN and one additional stage of CPM, and results are shown in Table II. For HCRN, parameters are introduced in the heatmap diffusion and refinement steps. For each CRU, in heatmap diffusion step, the parameter number of the deformation weight W_d^{l-1} is 4, and in heatmap refinement step, the parameter number of W_C^l and B_C^l are $1\times1\times2\times1$=2 and 1, respectively. Therefore, the total parameter number of HCRN including 14 CRUs for all body joints is $(4+(2+1))\times14$=98. Comparing with the large parameter number 6,563,712[1] of one stage CPM, the proposed HCRN is extremely lighter, further verifying its efficiency.

To more clearly see effectiveness of our proposed HCRN, we compare the performance of VGG16-HCRN with VGG16-holistic without HCRN. We can see that the performance is improved from 79.5% (without HCRN) to 81.3% (with HCRN), while the inference time is only slightly increased from 0.061 seconds to 0.067 seconds. This clearly shows the effectiveness and efficiency of the proposed CRU in extracting contextual information from easy joints to guide the localization of difficult ones. Besides, we can see that the proposed VGG16-HCRN model can achieve comparable results with the CPM model with two stages, while it only costs 1/3 time. Since the refinement process of CRU is only applied on the generated heatmaps, there are fewer parameters to learn and thus the refinement process is simple and fast. Indeed, CRU improves the performance on all joints. For instance, the improvement for head mainly comes from the head top which benefits from contextual provided by the neck. We observe improvements for

[1]For the parameter number of one stage to CPM, please refer to the network architecture in [12] for more details.

TABLE III
COMPARISON ON DIFFERENT SETTINGS OF CONFIDENCE MAP MODELS (BINARY AND GAUSSIAN) AND LOSS FUNCTIONS (CROSS ENTROPY AND L2) WITH THE PROPOSED HCRN ON LSP DATASET.

Methods	PCK
Binary Map + Cross Entropy Loss	81.3
Binary Map + L2 Loss	80.4
Gaussian Map + Cross Entropy Loss	79.8
Gaussian Map + L2 Loss	80.9

difficult joints are more significant, *e.g.*, for wrist and ankle from 68.7% to 72.1% and 69.9% to 73.6%, respectively.

HCRN can work well with advanced CNN architectures. From Table I, the proposed HCRN model consistently improves the performance of baselines Res101, CPM and HG models. In addition, we can observe that HCRN helps improve the accuracies for all kinds of body joints. Moreover, the improvements on difficult joints are obvious, for example, CPM-HCRN improves CPM on wrists and ankles from 74.7% to 75.9% and 80.3% to 82.0% PCK, respectively. Introducing HCRN only increases the inference time by 0.006 seconds. These results further demonstrate the effectiveness and efficiency of our HCRN model to refine the initial heatmaps produced from state-of-the-art network architectures via integrating and exploiting spatial contextual information from easy joints to assist estimations of difficult ones.

To verify the efficacy of the heatmap diffusion function defined in Eqn. (5) in CRU, we perform ablated experiments via directly concatenating the initial heatmap of one joint with the heatmap of its articulated neighbor joint in the previous layer. We evaluated both models with VGG16 and Res101 as the network backbone, denoted as VGG16-HCRN-w/o-Diffusion and Res101-HCRN-w/o-Diffusion, respectively, in Table I. We can see removing the diffusion step degrades the performance of both VGG16 and Res101 based models, from 81.3% to 80.2% and 84.2% to 83.3% PCK, respectively. In addition, the accuracies of all joints drop, which demonstrates the effectiveness of the heatmap diffusion step to spread contextual information of easy joints to guide the estimation of difficult ones. Moreover, we can also observe that both VGG16- and Res101-HCRN-w/o-Diffusion improve the performance of holistic baselines, further verifying benefits of the proposed hierarchical model via explicitly considering joint complexities for human pose estimation.

To illustrate the advantage of constructing binary map for body joints and utilizing cross entropy loss as supervision, we conduct experiments with VGG16-HCRN to compare with the commonly used Gaussian maps and L2 loss by state-of-the-arts models [12], [10] for pose estimation. Results are shown in Table III. We find that "Binary Map+L2 Loss" outperforms other settings, including "Gaussian Map+L2 Loss". The superiority of "Binary Map+L2 Loss" derives from the proposed hierarchical model individually estimates the location of each joint, which is better modeled as a binary classification problem than regression and easier to learn than Gaussian maps with soft labels.

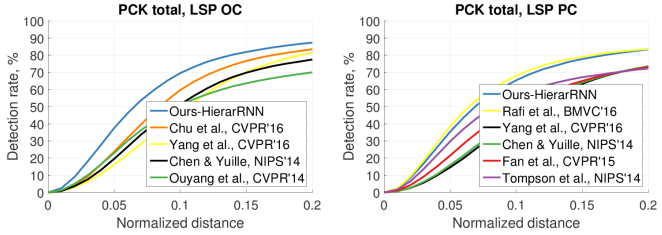
Comparison with State-of-the-arts We also compare our model with state-of-the-arts and show results in Table IV and Table V. From Table IV, we can see that Res101-HCRN already achieves highest 87.4% PCK on LSP dataset with Observer-Centric, significantly outperforming previous best [55] with 83.6% PCK. In addition, our model achieves the best results for all joints under PCK OC metrics

TABLE IV
EXPERIMENTS ON LSP DATASET WITH PCK OC METRIC. THE PERFORMANCE OF OUR PROPOSED MODEL RES101-HCRN AND STATE-OF-THE-ARTS ARE PRESENTED FOR COMPARISON.

Methods	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Avg
Res101-HCRN	**96.7**	**91.8**	**82.6**	**75.2**	**92.2**	**88.8**	**84.2**	**87.4**
Chu et al. [55]	93.7	87.2	78.2	73.8	88.2	83.0	80.9	83.6
Yang et al. [29]	90.6	89.1	80.3	73.5	85.5	82.8	68.8	81.5
Chen & Yuille [27]	91.5	84.7	70.3	63.2	82.7	78.1	72.0	77.5
Ouyang et al. [56]	86.5	78.2	61.7	49.3	76.9	70.0	67.6	70.0

TABLE V
EXPERIMENTS ON LSP DATASET WITH PCK PC METRIC. THE PERFORMANCE OF OUR PROPOSED MODELS RES101-, CPM- AND HG-HCRN, AND STATE-OF-THE-ARTS ARE PRESENTED FOR COMPARISON.

Methods	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Avg
HG-HCRN	**97.6**	**88.0**	**82.0**	**76.1**	**88.5**	**85.4**	**82.6**	**85.7**
CPM-HCRN	97.0	87.1	81.1	75.9	87.2	84.5	82.0	85.0
Res101-HCRN	96.8	86.8	79.7	75.8	85.7	84.8	80.1	84.2
Wei et al. [12]	96.9	86.7	80.4	74.7	86.7	83.3	80.3	84.1
Rafi et al. [57]	95.8	86.2	79.3	75.0	86.6	83.8	79.8	83.8
Yang et al. [29]	90.6	78.1	73.8	68.8	74.8	69.9	58.9	73.6
Chen & Yuille [27]	91.8	78.2	71.8	65.5	73.3	70.2	63.4	73.4
Carreira et al. [15]	90.5	81.8	65.8	59.8	81.6	70.6	62.0	73.1
Fan et al. [58]	92.4	75.2	65.3	64.0	75.7	68.3	70.4	73.0
Tompson et al. [11]	90.6	79.2	67.9	63.4	69.5	71.0	64.2	72.3



(a) (b)

Fig. 5. Pose estimation results over all PCK thresholds on LSP dataset. (a) is for Observer-Centric. (b) is for Person-Centric.

Table V shows the comparisons with state-of-the-arts trained with LSP and LSPET on PCK PC metric. We can see that HG-HCRN sets new state-of-the-art 85.7% PCK. Moreover, our single-stage model Res101-HCRN can achieve comparable performance 84.2% PCK with state-of-the-art 84.1% PCK. We also show the experimental results with efficient Res101-HCRN over all PCK thresholds in Fig. 5 to better illustrate its effectiveness. We compare Res101-HCRN to others with the best PCK performance under PC and OC settings, respectively. It can be seen that Res101-HCRN outperforms them w.r.t all thresholds under PCK OC metric on the LSP dataset, and achieves comparable results with state-of-the-arts under PCK PC metric. In Fig. 6, we show estimations for separate joints in different layers over all thresholds under PCK PC metric, demonstrating that the proposed model can achieve comparable performance with state-of-the-arts over all thresholds for all joints. In particular, for shoulder, knee and wrist, the proposed Res101-HCRN model even achieves slightly

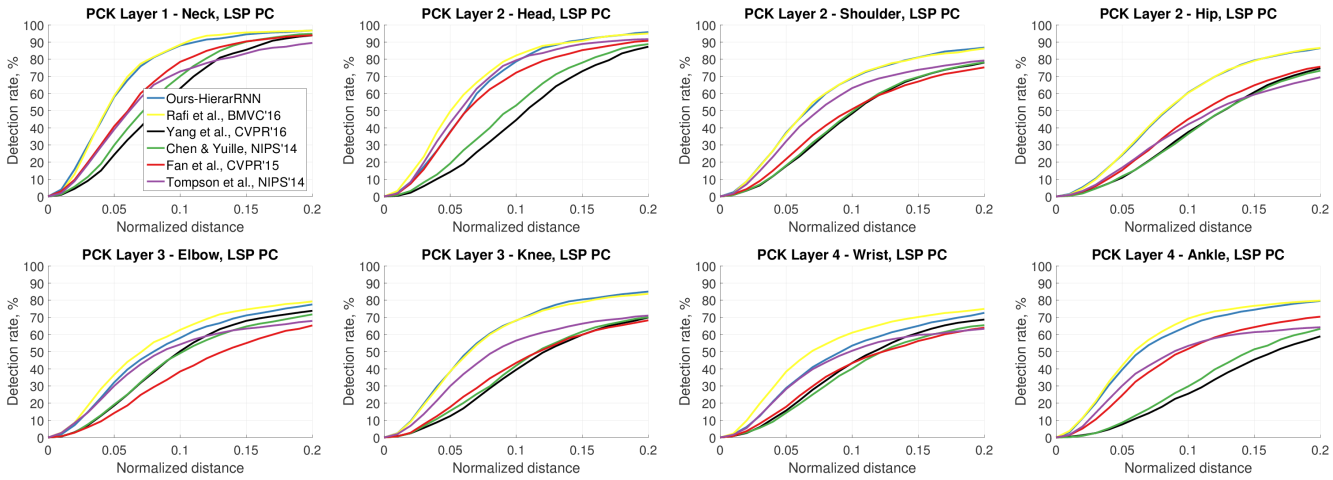


Fig. 6. Pose estimation results for layers in the proposed HCRN model over all PCK PC thresholds on LSP dataset.

TABLE VI
EXPERIMENTS ON THE LSP DATASET WITH PCP OC METRIC. THE PERFORMANCE OF OUR PROPOSED MODEL RES101-HCRN AND STATE-OF-THE-ARTS ARE PRESENTED FOR COMPARISON.

Methods	Torso	U.Leg	L.Leg	U.Arm	F.Arm	Head	Total
Res101-HCRN	**97.3**	**90.6**	**84.0**	**81.9**	65.1	**94.2**	**83.5**
Chu et al. [55]	95.4	87.6	83.3	76.9	**65.2**	89.6	81.1
Yang et al. [29]	96.5	88.7	81.7	78.8	66.7	83.1	81.1
Chen & Yuille [27]	92.7	82.9	77.0	69.2	55.4	87.8	75.0
Ouyang et al. [56]	88.6	77.8	71.9	61.9	45.4	84.3	68.7

TABLE VII
EXPERIMENTS ON THE LSP DATASET WITH PCP PC METRIC. THE PERFORMANCE OF OUR PROPOSED MODEL RES101-, CPM- AND HG-HCRN, AND STATE-OF-THE-ARTS ARE PRESENTED FOR COMPARISON.

Methods	Torso	U.Leg	L.Leg	U.Arm	F.Arm	Head	Total
HG-HCRN	**98.2**	86.8	**82.5**	**78.9**	**68.7**	**95.0**	**82.4**
CPM-HCRN	97.3	86.2	81.8	78.1	67.1	94.4	81.7
Res101-HCRN	96.8	86.5	80.7	78.8	67.8	93.7	81.8
Wei et al. [12]	97.1	85.7	81.0	77.8	66.4	94.1	81.3
Rafi et al. [57]	97.6	**87.3**	80.2	76.8	66.3	93.3	81.2
Yang et al. [29]	95.6	78.5	71.8	72.3	61.8	83.9	74.8
Chen & Yuille [27]	96.0	77.3	72.2	69.7	58.1	85.6	73.6
Fan et al. [58]	95.4	77.7	69.8	62.8	49.1	86.6	70.1
Tompson et al. [11]	90.3	70.4	61.1	63.0	51.3	83.7	66.6

better results than state-of-the-arts, although its architecture is simpler and it is more efficient.

To further prove the effectiveness of our model, we also evaluate it under the PCP metric and compare with state-of-the-arts. Experimental results are summarized in Table VI and Table VII. From Table VI, we can see that Res101-HCRN achieves the best overall performance 83.5% on the LSP dataset under PCP OC metric. We can also observe that our model achieves the highest scores on all joints except for the forearm. From Table VII, HG-HCRN sets new state-of-the-art 82.4% under PCP PC metric, in addition, it also achieves the best performance for most of joints, except for under legs.

Qualitative Results Qualitative results on the LSP dataset are given in the top two rows of Fig. 7. We visualize the localization results of body joints according to the complexity-aware hierarchy, where joints in the same layer are annotated with the same color. From Fig. 7, we can see our model can deal with various highly articulated poses, variant orientations, and overlapping people. It is also robust to the cluttered background. For example, for the 10th image in the second row, our model can provide correct pose estimation in presence of the cluttered tree in the background.

TABLE VIII
EXPERIMENTS ON THE FILC DATASET WITH PCK OC METRIC. THE PERFORMANCE OF OUR PROPOSED MODEL RES101-, CPM- AND HG-HCRN, AND STATE-OF-THE-ARTS ARE PRESENTED FOR COMPARISON.

Methods	Elbow	Wrist
HG-HCRN	**99.2**	**97.3**
HG [10]	99.0	97.0
CPM-HCRN	98.1	95.7
CPM [12]	97.6	95.0
Res101-HCRN	97.4	95.1
Res101-Holistic	95.5	93.0
Chen et al. NIPS'14 [27]	95.3	92.4
Tompson et al. CVPR'15 [59]	93.1	89.0
Toshev et al., CVPR'14 [60]	92.3	82.0
Sapp et al., CVPR'13 [18]	76.5	59.1

B. Experiments on FLIC Dataset

Frames Labeled In Cinema (FLIC) [18] dataset contains 5,003 images extracted from popular Hollywood movies by a person detector. Among these images, 3,987 images are used for training and 1,016 images for testing. Different from LSP, FLIC dataset only provides 10 upper body joint annotations. Along with joint locations, torso boxes are also provided in FLIC dataset. We crop each person from the image based on the torso box by $3\times$ enlargement. We use similar data augmentation strategies adopted on LSP dataset to augment the training samples. Images are also resized and padded into 368×368 as input to CNN for both training and testing. Target heatmaps are generated similarly as in LSP only for the annotated joints and background. For training and testing on FLIC, our approach only constructs hierarchies for annotated joints, which also shows its flexibility by stacking CRUs for different structures. We use PCK@0.2 with Observer-Centric to evaluate our approach and compare with state-of-the-arts.

Fig. 7. Qualitative results on LSP dataset (top two rows) and FLIC dataset (the bottom row). Our model can provide accurate and robust pose estimation even in some challenging conditions, *e.g.*, cluttered background (the bottom row), extreme poses (the top-left result).

TABLE IX
EXPERIMENTS ON THE MPII DATASET WITH PCKH@0.5 METRIC. THE PERFORMANCE OF OUR PROPOSED MODEL HG-HCRN AND STATE-OF-THE-ARTS ARE PRESENTED FOR COMPARISON.

Methods	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	PCKh
HG-HCRN	**98.4**	**96.3**	**91.7**	**87.7**	**90.6**	**87.4**	83.3	**91.2**
Newell et al. [10]	98.2	96.3	91.2	87.1	90.1	87.4	**83.6**	90.9
Bulat & Tzimiropoulos [61]	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7
Wei et al. [12]	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5
Insafutdinov et al. [62]	96.8	95.2	89.3	84.4	88.4	83.4	78.0	88.5
Rafi et al. [57]	97.2	93.9	86.4	81.3	86.8	80.6	73.4	86.3
Gkioxari et al. [16]	96.2	93.1	86.7	82.1	85.2	81.4	74.1	86.1
Hu and Ramanan [63]	95.0	91.6	83.0	76.6	81.9	74.5	69.5	82.4
Tompson et al. [59]	96.1	91.9	83.9	77.8	80.9	72.3	64.8	82.0
Carreira et al. [15]	95.7	91.7	81.7	72.4	82.8	73.2	66.4	81.3
Tompson et al. [11]	95.8	90.3	80.5	74.3	77.6	69.7	62.8	79.6
Pishchulin et al. [64]	74.3	49.0	40.8	34.1	36.5	34.4	35.2	44.1

Experimental results are shown in Table VIII. We can see HCRN improves all baseline models Res101, CPM and HG for both elbows and wrists, further proving the effectiveness of the proposed CRUs in utilizing contextual information from difficulty joints. Compared with state-of-the-art methods, HG-HCRN achieves new state-of-the-art 99.2% and 97.3% PCK for the elbow and wrist, respectively.

Qualitative results on FLIC are shown in the bottom row of Fig. 7, in which we only visualize the upper body joints predicted by our model. It can be seen that our approach is able to provide accurate and robust pose estimation, even in challenging conditions.

C. Experiments on MPII Dataset

MPII Human Pose Single-Person (MPII) dataset [19] is a state-of-the-art benchmark for evaluating human pose estimation algorithms. It contains 19,185 training and 7,247 testing images in every day human activities, and each people is annotated with 16 joints. In addition, the scale and center of each people are also provided. We utilize the provided scale to resize the training samples to roughly the same size. Then, we crop each sample from the original image around the center position. We resize and pad the cropped samples to 384×384 as input for training our model on MPII dataset. We also augment the training samples by random scaling in $[0, 7, 1.3]$, rotation in $[-40°, 40°]$ and flipping. The target heatmaps are constructed by assigning a positive label 1 at each location within 12 pixels to the ground truth, and otherwise a negative label 0. We use the same strategy adopted in [59] to split the validation set consisting of 2,958 images for supervising our training process. We utilize our best model HG-HCRN to compare with state-of-the-arts on MPII dataset. The evaluation is based on the official PCKh metric [19], following conventions.

The experimental results are shown in Table IX. We can see that the HCRN model improves the performance of baseline Hourglass network [10] from 90.9% PCKh to 91.2% PCKh, achieving new state-of-the-art on MPII dataset. In addition, HCRN improves the performance of most of the joints, except the ankle. The reason lies in our re-implementation of Hourglass network can not achieve same performance of ankles with the original one. These results evaluate the effectiveness of HCRN for refining pose estimation results, even of the state-of-the-art Hourglass model, via considering joint complexities.

Qualitative results on MPII dataset are shown in Fig. 8. We can observe that our proposed model can deal with various challenging scenarios, *e.g.*, occlusions (1st example of 1st row, last example of 2nd row, and examples of 3rd row), cluttered background (4th row), and large pose variations (last two rows). These results further verify the effectiveness of our proposed HCRN model.

D. Experiments on MSCOCO dataset

To further verify the efficacy of our HCRN model, we conduct experiments on the multi-person pose estimation benchmark MSCOCO [20], which contains about 60,000 training images with 17 annotated body joints per person.

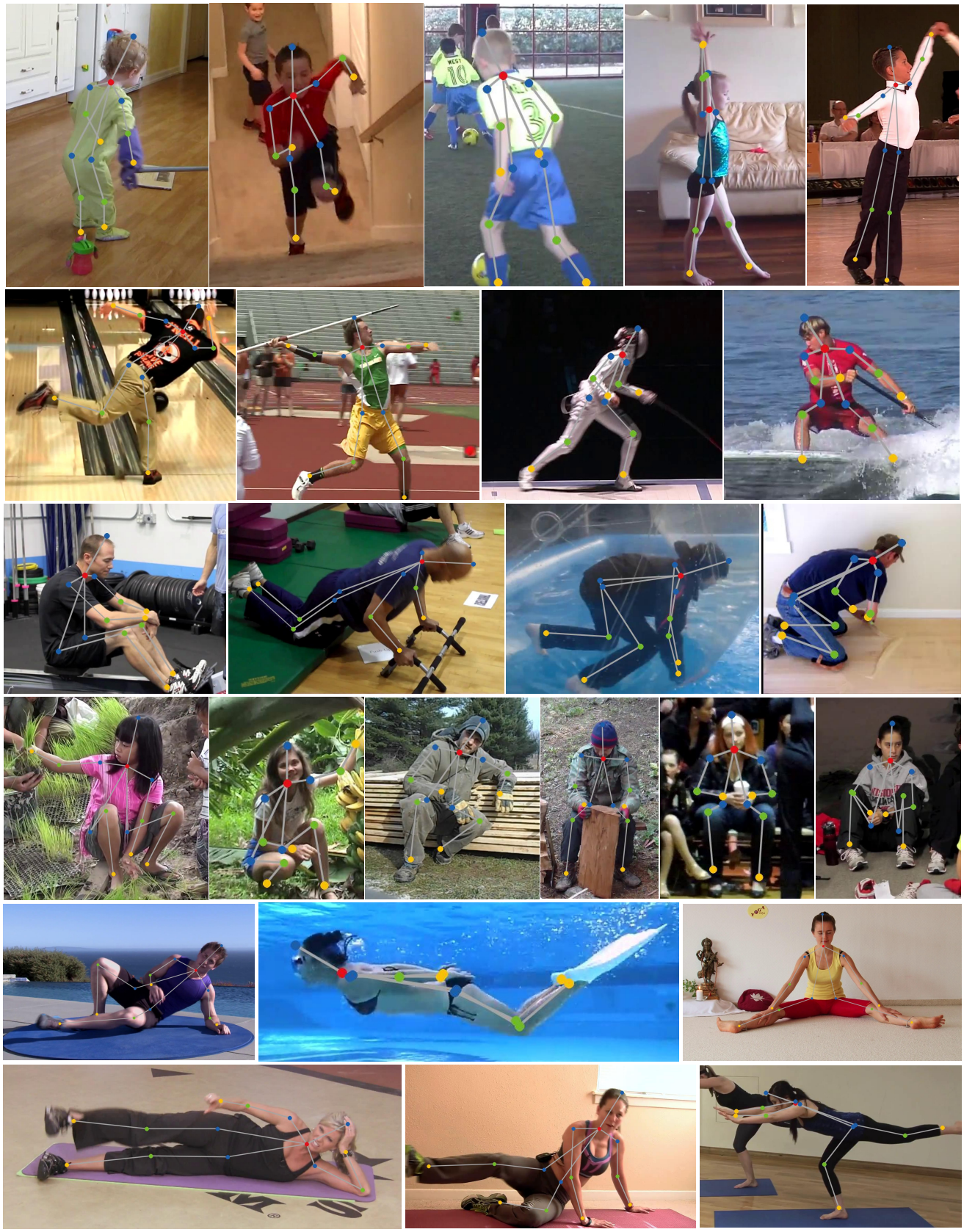


Fig. 8. Qualitative results on MPII dataset. The proposed HCRN model can deal with various challenging scenarios, *e.g.*, occlusions (1st example of 1st row, last example of 2nd row, and examples of 3rd row), cluttered background (4th row), and large pose variations (last two rows).

TABLE X
COMPARISON WITH STATE-OF-THE-ARTS ON MSCOCO TEST-DEV.

Methods	AP	Times[s]
OpenPose [65]	0.618	1.24
OpenPose + HG-HCRN	0.625	2.00
MaskRCNN [66]	0.627	0.190
MaskRCNN + HG-HCRN	0.630	0.960

Evaluations are conducted on the test-dev subset, including roughly 20,000 images, with the official Average Precision (AP) as metric. Our model is designed for the *single-person* pose estimation task, not for *multi-person* tasks. To extend the model to the *multi-person* case for comparison with state-of-the-arts, we first generate person detections via either ResNet-50-FPN based MaskRCNN[2] or tight bounding boxes covering individual joints from multi-person pose estimation results of OpenPose[3]. Then we perform single-person pose estimation using the proposed HG-HCRN model for each person detection individually. Here, HG-HCRN is followed the same setting as on MPII, *i.e.* using 8-stack Hourglass network as backbone, resizing and padding the cropped images as 384×384 as input, and performing 3-scale testing. Results are shown in Tab. X.

We can observe that the proposed HG-HCRN improves OpenPose from 61.8% AP to 62.5% AP. In addition, "MaskRCNN + HG-HCRN" achieves 63.0% AP, outperforming MaskRCNN (62.7% AP) and setting new state-of-the-art. The results clearly demonstrate effectiveness of our HCRN model for human pose estimation

In terms of speed, "MaskRCNN + HG-HCRN" costs 0.960s for processing one image. It is faster than OpenPose (1.24s/image) though slower than MaskRCNN due to the extra HG-HCRN component. In addition, its speed will linearly decrease when the number of person in a image increases, since it follows the *top-down* strategy and needs to run HG-HCRN for each person bounding box sequentially.

V. CONCLUSION

In this paper, we propose the Hierarchical Contextual Refinement Networks (HCRNs) for effectively and efficiently predicting human body joints from easy to difficult. The proposed method divides body joints into four layers to build the joint hierarchy. We also develop a novel Contextual Refinement Unit (CRU) which enables auto-diffusion of joint detections results to effectively transfer informative context from easy joints to difficult ones and is organized according to the hierarchy to transfer informative context from easy joints to help localize difficult joints. Combined with a front-end CNN, the unified framework is end-to-end trainable. Evaluation of the HCRN model on multiple challenging benchmarks demonstrates HCRN achieves superior performance over both holistic and multi-stage approaches and offers advantageous efficiency. In addition, the proposed model also achieves new state-of-the-art on multiple benchmarks.

[2]We use the official codes and pre-trained models from the "Detectron" repository implemented with caffe2 in the following link: https://github.com/facebookresearch/Detectron

[3]We use the codes and pre-trained models implemented with caffe in the following link: https://github.com/CMU-Perceptual-Computing-Lab/openpose

ACKNOWLEDGMENT

Jiashi Feng was partially supported by NUS IDS R-263-000-C67-646, ECRA R-263-000-C87-133 and MOE Tier-II R-263-000-D17-112.

REFERENCES

[1] H. Y. Lin and T. W. Chen, "Augmented reality with human body interaction based on monocular 3d pose estimation," in *ACIVS*, 2010.
[2] S. Obdržálek, G. Kurillo, J. Han, T. Abresch, and R. Bajcsy, "Real-time human pose detection and tracking for tele-rehabilitation in virtual reality," *Studies in Health Technology and Informatics*, vol. 173, pp. 320–324, 2012.
[3] S.-R. Ke, L. Zhu, J.-N. Hwang, H.-I. Pai, K.-M. Lan, and C.-P. Liao, "Real-time 3d human pose estimation from monocular view with applications to event detection and video gaming," in *AVSS*, 2010.
[4] D. Zhang and M. Shah, "Human pose estimation in videos," in *ICCV*, 2015.
[5] M. Andriluka, S. Roth, and B. Schiele, "Monocular 3d pose estimation and tracking by detection," in *CVPR*, 2010.
[6] C. Wang, Y. Wang, and A. L. Yuille, "An approach to pose-based action recognition," in *CVPR*, 2013.
[7] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Parsing clothing in fashion photographs," in *CVPR*, 2012.
[8] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, "Progressive search space reduction for human pose estimation," in *CVPR*, 2008.
[9] A. Haque, B. Peng, Z. Luo, A. Alahi, S. Yeung, and L. Fei-Fei, "Towards viewpoint invariant 3d human pose estimation," in *ECCV*, 2016.
[10] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *ECCV*, 2016.
[11] 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.
[12] S. E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *CVPR*, 2016.
[13] K. M. Knutzen, *Kinematics of human motion*. Wiley Online Library, 1998.
[14] X. Fan, K. Zheng, Y. Zhou, and S. Wang, "Pose locality constrained representation for 3d human pose reconstruction," in *ECCV*, 2014.
[15] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, "Human pose estimation with iterative error feedback," in *CVPR*, 2016.
[16] G. Gkioxari, A. Toshev, and N. Jaitly, "Chained predictions using convolutional neural networks," in *ECCV*, 2016.
[17] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *BMVC*, 2015.
[18] B. Sapp and B. Taskar, "Modec: Multimodal decomposable models for human pose estimation," in *CVPR*, 2013.
[19] M. Andriluka, P. Leonid, G. Peter, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *CVPR*, 2014.
[20] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014.
[21] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 55–79, 2005.
[22] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *CVPR*, 2009.
[23] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Poselet conditioned pictorial structures," in *CVPR*, 2013.
[24] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *CVPR*, 2011.
[25] M. W. Lee and I. Cohen, "Proposal maps driven mcmc for estimating human body pose in static images," in *CVPR*, 2004.
[26] S. Ioffe and D. A. Forsyth, "Probabilistic methods for finding people," *International Journal of Computer Vision*, vol. 43, no. 1, pp. 45–68, 2001.
[27] X. Chen and A. L. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," in *NIPS*, 2014.
[28] 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.
[29] 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.
[30] A. Bulat and G. Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression," in *ECCV*, 2016.

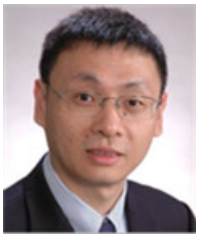
[31] V. Belagiannis and A. Zisserman, "Recurrent human pose estimation," in *FG*, 2017.
[32] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *CVPR*, 2017.
[33] Y. Chen, C. Shen, X.-S. Wei, L. Liu, and J. Yang, "Adversarial posenet: A structure-aware convolutional network for human pose estimation," in *ICCV*, 2017.
[34] Z. Liu, S. Yan, P. Luo, X. Wang, and X. Tang, "Fashion landmark detection in the wild," in *ECCV*, 2016.
[35] A. Graves, A. R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *ICASSP*, 2013.
[36] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[37] B. X. Nie, P. Wei, and S.-C. Zhu, "Monocular 3d human pose estimation by predicting depth on joints," in *ICCV*, 2017.
[38] A. R. Zamir, T.-L. Wu, L. Sun, W. B. Shen, B. E. Shi, J. Malik, and S. Savarese, "Feedback networks," in *CVPR*, 2017.
[39] V. Belagiannis, C. Rupprecht, G. Carneiro, and N. Navab, "Robust optimization for deep regression," in *ICCV*, 2015.
[40] M. Sun and S. Savarese, "Articulated part-based model for joint object detection and pose estimation," in *ICCV*, 2011.
[41] Y. Tian, C. L. Zitnick, and S. G. Narasimhan, "Exploring the spatial hierarchy of mixture models for human pose estimation," in *ECCV*, 2012.
[42] C. Ionescu, J. Carreira, and C. Sminchisescu, "Iterated second-order label sensitive pooling for 3d human pose estimation," in *CVPR*, 2014.
[43] X. Sun, Y. Wei, S. Liang, X. Tang, and J. Sun, "Cascaded hand pose regression," in *CVPR*, 2015.
[44] D. Tang, J. Taylor, P. Kohli, C. Keskin, T. K. Kim, and J. Shotton, "Opening the black box: Hierarchical sampling optimization for estimating human hand pose," in *ICCV*, 2015.
[45] Q. Ye, S. Yuan, and T. K. Kim, "Spatial attention deep net with partial pso for hierarchical hybrid hand pose estimation," in *ECCV*, 2016.
[46] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM international conference on Multimedia*, 2014.
[47] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *CoRR*, 2015.
[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
[49] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *ICLR*, 2015.
[50] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural Networks for Machine Learning*, 2012.
[51] S. Johnson and M. Everingham, "Learning effective human pose estimation from inaccurate annotation," in *CVPR*, 2011.
[52] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 12, pp. 2878–2890, 2013.
[53] M. Eichner and V. Ferrari, "Appearance sharing for collective human pose estimation," in *ACCV*, 2012.
[54] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *BMVC*, 2010.
[55] X. Chu, W. Ouyang, H. Li, and X. Wang, "Structured feature learning for pose estimation," in *CVPR*, 2016.
[56] W. Ouyang, X. Chu, and X. Wang, "Multi-source deep learning for human pose estimation," in *CVPR*, 2014.
[57] U. Rafi, I. Kostrikov, J. Gall, and B. Leibe, "An efficient convolutional network for human pose estimation," in *BMVC*, 2016.
[58] 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.
[59] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *CVPR*, 2015.
[60] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *CVPR*, 2014.
[61] A. Bulat and G. Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression," in *ECCV*, 2016.
[62] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," in *ECCV*, 2016.
[63] P. Hu and D. Ramanan, "Bottom-up and top-down reasoning with hierarchical rectified gaussians," in *CVPR*, 2016.
[64] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Strong appearance and expressive spatial models for human pose estimation," in *CVPR*, 2013.
[65] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.
[66] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *ICCV*, 2017.

Xuecheng Nie received his B.S. and M.Eng. degrees in School of Computer Software from Tianjin University, Tianjin, China, in 2012 and 2015, respectively. He is currently a Ph.D. candidate at Learning and Vision Lab, Department of Electrical and Computer Engineering, National University of Singapore, Singapore. His research interests focus on Computer Vision, Deep Learning, specially at Human Pose Estimation.

Jiashi Feng received the Ph.D. degree from the National University of Singapore (NUS) in 2014. He was a Post-Doctoral Research Follow with the University of California, Berkeley. He joined NUS as a Faculty Member, where he is currently an Assistant Professor with the Department of Electrical and Computer Engineering. His research areas include computer vision, machine learning, object recognition, detection, segmentation, robust learning and deep learning.

Junliang Xing received his dual B.S. degrees in computer science and mathematics from Xi'an Jiaotong University, Shaanxi, China, in 2007, and the Ph.D. degree in computer science from Tsinghua University, Beijing, China, in 2012. He is currently an Associate Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. Dr. Xing was the recipient of Google Ph.D. Fellowship 2011, the Excellent Student Scholarships at Xi'an Jiaotong University from 2004 to 2007 and at Tsinghua University from 2009 to 2011. He has published more than 70 papers on international journals and conferences. His current research interests mainly focus on computer vision problems related to faces and humans.

Shengtao Xiao is currently a Research Scientist in Artificial Intelligence Institute at Qihoo 360 International. He received his Bachelor and Doctor Degree from National University of Singapore (NUS) in 2013 and 2018 respectively. His research interest includes Landmark Detection, Pattern Classification, and Facial Image Analysis.



Shuicheng Yan is currently the Vice-President and the Chief Scientist with Qihoo 360 Technology Company Ltd., and the Head of the 360 Artificial Intelligence Institute. He is also a tenured Associate Professor with the National University of Singapore. He has authored/co-authored over 500 high quality technical papers, with Google Scholar citation over 25 000 times and an h-index 70. His research areas include computer vision, machine learning, and multimedia analysis. He is an IAPR Fellow and the ACM Distinguished Scientist. His team received seven times winner or honorable-mention prizes in five years over PASCAL, VOC, and ILSVRC competitions, which are core competitions in the field of computer vision, along with over ten times the Best (student) Paper Awards and especially a Grand Slam with the ACM MM, the top conference in the field of multimedia, including the Best Paper Award, the Best Student Paper Award, and the Best Demo Award. He is a TR Highly Cited Researcher of 2014, 2015, and 2016.