IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 7, 2019, accepted February 23, 2019, date of publication February 28, 2019, date of current version March 20, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2902330

Efficient Human Pose Estimation in Hierarchical Context

FENG ZHANG[1], XIATIAN ZHU[2], AND MAO YE[1]

[1]Key Laboratory for NeuroInformation, Ministry of Education, Center for Robotics, School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[2]Vision Semantics Limited, London E1 4NS, U.K.

Corresponding author: Mao Ye (cvlab.uestc@gmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61773093, in part by the National Key R&D Program of China (Intelligent Processing Technology of Multi-source Litigation Letters and Visits National) under Grant 2018YFC0831800, in part by the Important Science and Technology Innovation Projects in Chengdu under Grant 2018-YF08-00039-GX, and in part by the Research Programs of the Sichuan Science and Technology Department under Grant 2016JY0088 and Grant 17ZDYF3184.

ABSTRACT Most existing human pose estimation methods focus on enhancing the accuracy performance alone while ignoring the critical model efficiency issue. This dramatically limits their scalability and deployability in large-scale applications. In this paper, we consider the under-studied model efficiency problem in pose estimation. We demonstrate the advantages and potential of hierarchical context learning in the convolutional neural network. Specifically, we formulate a novel *hierarchical context network* (HCN) architecture that can be trained and deployed efficiently while achieving competitive model generalization capability. This is achieved by progressively forming and imposing multi-granularity context information during the pose regression learning process in a coarse-to-fine manner. The extensive comparative evaluations validate the superiority of the proposed HCN over a wide variety of the state-of-the-art human pose estimation models on two challenging benchmarks: MPII and LSP.

INDEX TERMS Fast deployment, human pose estimation, hierarchical context, model cost-effectiveness.

I. INTRODUCTION

Human pose estimation is a task of identifying local body parts/joints in scene images [1]. It is intrinsically challenging due to the unconstrained covariates in body appearance, viewpoint, illumination, occlusion, and background clutter. Earlier methods [2], [3], [34], [46] rely on hand-crafted representations (e.g. SIFT, HoG) and shallow recognition models (e.g. pictorial structures) learned independently, hence often yielding suboptimal performance. Deep CNN methods dominate the recent progress by jointly learning more discriminative features and inference models [9], [11], [41]. However, existing deep models are typically with complex designs therefore sacrificing the model efficiency in training and test, i.e. poor cost-effectiveness (Fig 1). This significantly limits their scalability and usability in real-world large scale deployments. Whilst the accuracy has clearly improved, how to design efficient deep models remains under-studied and challenging.

In this work, we investigate the pose estimation efficiency problem. Our **contributions** are summarized as below:

1) We study the model efficiency issue in pose estimation for both training and test, which is largely neglected in the literature but critical in scaling up real-world deployments.
2) We formulate a novel *Hierarchical Context Network* (HCN) framework capable of being trained and deployed efficiently while simultaneously achieving competitive performance. This is inspired by the coarse-to-fine human visual perception principle. Specifically, HCN is designed to progressively integrate multi-granularity context constraints into the pose regression learning in a coarse-to-fine sequential manner so that the model size and optimization search space can be effectively minimized without sacrificing the model discrimination capability.
3) To validate the effectiveness, we further implement a concrete HCN pose estimation model by exploiting the Hourglass CNN module [26] and introducing extra design enhancements.

The associate editor coordinating the review of this manuscript and approving it for publication was Shuai Liu.

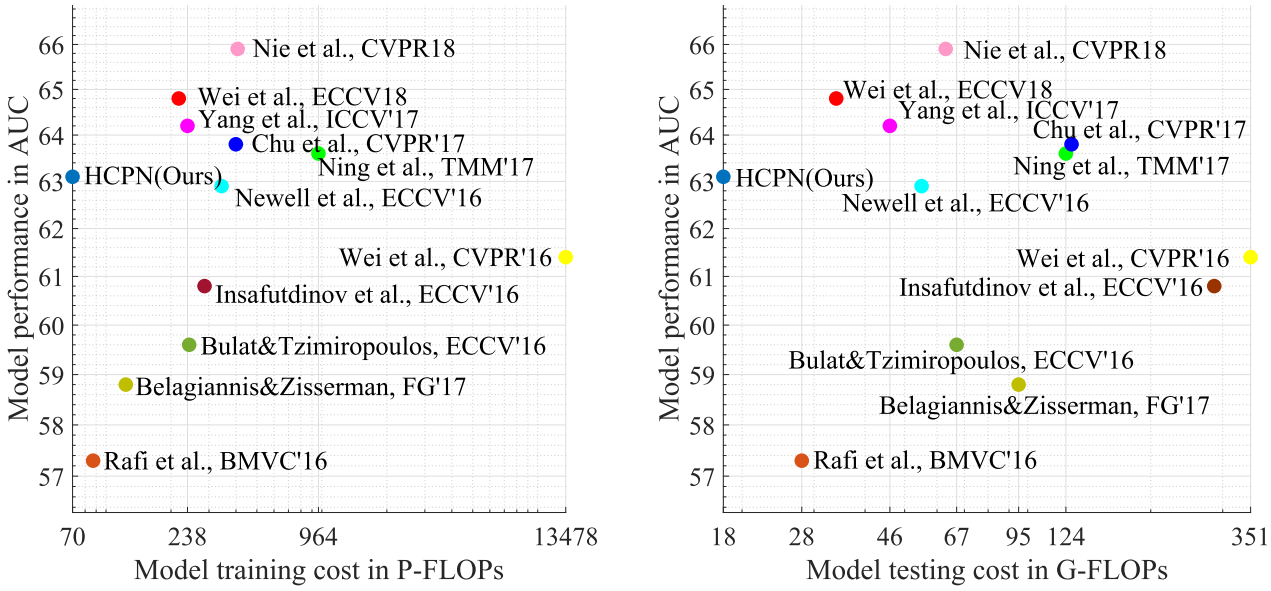


FIGURE 1. Comparison of state-of-the-art human pose estimation methods in model (left) training and (right) test costs. Unit: 10^{15} (P) or 10^{9} (G) FLOPs (FLoating point OPerations).

Extensive comparative evaluations demonstrate the performance superiority and advantages of the proposed HCN approach over a wide range of state-of-the-art human pose models in terms of trade-off between model efficiency and accuracy performance on two challenging benchmarking datasets: MPII [1], and Leeds Sports Pose (LSP) [22]. The source codes will be publicly released.

II. RELATED WORK

The research attempts on human pose estimation have gradually shifted from hand-crafted feature based approaches [2], [3], [10], [30], [34], [46] to deep learning paradigm since "DeepPose" [41]. Beyond performing the direct joint location regression [41], Tompson *et al.* [39] adopt a multi-resolution sliding window strategy in a Siamese network [5] to refine the locations. Some works additionally integrate the spatial relationships between joints [10], [15], [31].

Another common approach is making successive predictions by stacked inference [9], [18], [26], [43]. Whilst significant accuracy gains have been generated, these existing methods overlooked the critical model efficiency issue as studied in this work.

There have been a few attempts at devising efficient pose models. In particular, Bulat and Tzimiropoulos [7] develop binarised nets for better test efficiency but sacrifices significantly the accuracy. Rafi et al. [32] boost the model training efficiency by using multi-scale learning and multi tricks including optimized learning rate and batch normalization. Cao et al. [8] seek for system-level real-time performance of multi-person pose estimation via joint part detection and association. In contrast, we present a unified network architecture specialized for improving the model training and test efficiency whilst simultaneously retaining the model generalization capability. Recently, more efficient U-Net [37], [38] and low-bit quantized nets have been concurrently developed. Conceptually, our method is largely orthogonal to the existing approaches in design from a different learning aspect.

III. MODEL DESIGN

Problem Formulation: To train a pose model, we often have a set of N labeled training samples $\mathcal{D}=\{\boldsymbol{I}_i, Z_i\}_{i=1}^{N}$, where $Z_i=\{z_1, \cdots, z_J\}$ (with $z=(u,v)\in\mathbb{R}^2$) defines the ground-truth locations of all J joints of the training image $\boldsymbol{I}_i \in \mathbb{R}^{h\times w\times 3}$ (h and w the image height and width). We aim to formulate a deep learning pose estimator, $f(\cdot;\ \boldsymbol{\theta})$, optimized to identify these joints in form of generating J 2D spatial confidence maps $\{\boldsymbol{m}_1, \cdots, \boldsymbol{m}_J\}$ with size $h\times w$ for an input image. Each map corresponds to an individual joint. We formally express this process as:

$$f(\boldsymbol{I};\boldsymbol{\theta}) = \{\boldsymbol{m}_i\}_{i=1}^{J}, \quad \boldsymbol{m}_i \in [0,1]^{h\times w} \tag{1}$$

where $\boldsymbol{m}_i(u,v)$ specifies the model inference confidence score of assigning the location (u,v) to the i-th joint.

In contrast to existing methods stacking multiple identical blocks [26] to gain higher modeling capacity which is expensive, we design a new CNN architecture offering *fast trainable* and *deployable* capabilities characterized by a *hierarchical context learning* mechanism with only little model generalization degradation.

A. HIERARCHICAL CONTEXT NETWORK

Design Rational: To formulate an efficient pose model, we explore the idea of *hierarchical context learning*, moti-

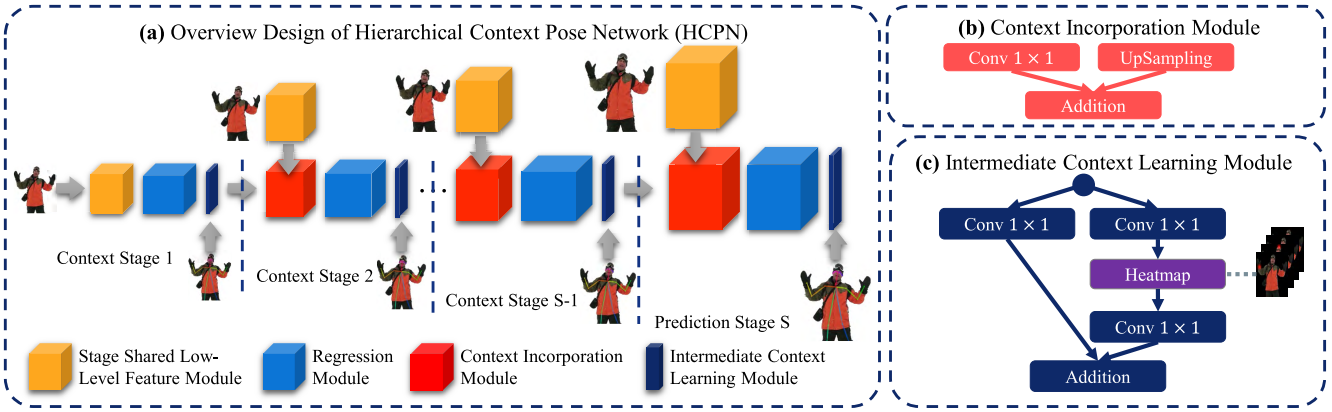


FIGURE 2. **Overview of the Hierarchical Context Network (HCN). The HCN consists of S stages ($S-1$ context stages and 1 prediction stage). All stages share the same structure with a low level feature module, a context incorporation module, a pose backbone module, and an intermediate context learning module. The first stage has no context incorporation module. The last stage replaces the intermediate context learning module with a pose prediction module (a 1×1 conv layer).**

vated by the psychophysical research that human visual perception can leverage jointly both global contextual and local saliency information in object detection and recognition [24], [33]. Specifically, the contextual clues allow the perception system to narrow down the searching space *efficiently* and *effectively* in a coarse-to-fine manner. We hypothesize that, such a principle may offer a means of establishing more efficient pose estimation models with satisfactory performance.

Overview: In light of this inspiration, we formulate a *Hierarchical Context Network* (HCN) architecture (see overview in Fig 2(a)). The HCN involves two types of stages: (1) $S-1$ *context* stages and (2) one *prediction* stage. Any context stage takes as input a small resolution of the same image $\boldsymbol{I}$ to learn the corresponding granularity of context. The context can be learned *efficiently*, because simple features suffice to represent it [24]. It is this property that opens up the possibility of boosting the model efficiency whilst maintaining the performance – *the foundation of designing HCN*. In comparison, the prediction stage takes the finest input (i.e. highest resolution), yielding the final model predictions. Stacking all stages together forms a feed-forward HCN with increasingly fine-grained input resolutions, enabling to sequentially learn *hierarchical context* information for efficient pose inference. Critically, HCN is flexible in integrating distinct CNN module designs therefore providing a generic solution with the potential of benefiting more advanced block designs. We provide more details of key HCN components below.

1) CONTEXT INCORPORATION

For sequentially embedding the context information to form a multi-level semantic representation, we need an incorporation mechanism. We achieve this by carrying the context output of the preceding stage on to the current context or prediction stage with a ***Context Incorporation*** (CI) module (Fig 2(b)). Specifically, the input to a CI module in the s-th stage includes: (1) the higher-level context $\boldsymbol{x}_{(s-1)}^{\text{ctx}}$ from the preceding stage (Fig 2(c)) and (2) the low-level features $\boldsymbol{x}_s^{\text{llv}}$ from the current stage, both of which are used to induce a context enriched representation $\boldsymbol{x}_s^{\text{input}}$ as the s-th stage's input. Formally, we formulate this integration by a learnable function $g_s^{\text{ctx}}(\cdot; \boldsymbol{\gamma}_s)$ as:

$$\boldsymbol{x}_s^{\text{input}} = g_s^{\text{ctx}}\left(\boldsymbol{x}_{(s-1)}^{\text{ctx}}, \boldsymbol{x}_s^{\text{llv}}; \boldsymbol{\gamma}_s\right) \tag{2}$$

where $\boldsymbol{\gamma}_s$ is the parameters of the CI module at the s-th stage. Given the semantic gap between $\boldsymbol{x}_{(s-1)}^{\text{ctx}}$ and $\boldsymbol{x}_s^{\text{llv}}$, we add a 1×1 sized conv layer to $\boldsymbol{x}_s^{\text{llv}}$. This helps to improve the fusion compatibility. We perform bi-linear upsampling on $\boldsymbol{x}_{(s-1)}^{\text{ctx}}$ to match the spatial size of $\boldsymbol{x}_s^{\text{llv}}$ in prior to merging them by element-wise addition.

Discussion: The context incorporation takes place between every two adjacent stages. In doing so, *hierarchical context learning* can be naturally and progressively realized in every mini-batch training. We do not use the Batch Normalization (BN) [21] for context incorporation, because the two input signals carry respective scale information at distinct semantics levels which can be ruined by BN. We will verify this design (Table 5).

2) INTERMEDIATE CONTEXT LEARNING SUPERVISION

We exploit the intermediate supervision mechanism in HCN (Fig 2(c)), inspired by previous pose models [13], [26] and cross-layer skip design [19] in the sense of enhancing the accessibility of ground-truth information in training. However, the objective of our model significantly differs as the supervision across HCN stages aims to enhance contextual constraint learning, instead of predicting the output pose structures.

Specifically, we pose a supervised loss constraint based on the ground-truth Z at the s-th context stage:

$$\mathcal{L}_s^{\text{ctx}} = g_{\text{loss}}\left(M_s, Z, s\right), \text{ with } M_s = h_s^{\text{ctx}}(\boldsymbol{x}_s^{\text{input}}; \boldsymbol{\theta}_s) \tag{3}$$

where $g_{\text{loss}}(\cdot)$ defines a loss function, and M_s indicates joint predictions (the confidence maps in Fig 2(c)).

Discussion: This multi-supervision scheme not only addresses the notorious difficulty of vanishing gradients but also makes the context information stronger and more relevant. Crucially, this design matches our ultimate objective of boosting model efficiency through reducing the parameter size without performance drop.

3) STAGE SHARED LOW-LEVEL FEATURES

We construct all stages on a *shared* low-level feature (LLF) module (Fig 2(a)) to facilitate cross-stage commonality learning. The intuition is that, the starting layers capture elementary patterns such as edges and corners commonly useful at resolutions. This is in a spirit of multi-task learning [14].

In particular, we rescale a raw image $\boldsymbol{I}$ to generate S samples $\{\boldsymbol{I}^s\}_{s=1}^S$, each of which matches the resolution of a specific stage. For the s-the stage, we feed $\boldsymbol{I}^s$ into the shared LLF module and obtain the corresponding representation as:

$$\boldsymbol{x}_s^{\text{llv}} = f_{\text{llv}}(\boldsymbol{I}^s; \boldsymbol{\eta}), \tag{4}$$

where $\boldsymbol{\eta}$ represents the module parameters. The LLF module is applicable to different input resolutions with the output feature maps linearly proportional to the input size.

Discussion: Sharing low-level layers reduces the model parameter size and hence model overfitting risks, making the model further concise and efficient. This is also useful to make pose models robust against unconstrained visual ambiguity especially when the labeled training data is small.

4) LOSS FUNCTION

For the model training, we adopt the Mean-Squared Error (MSE) as the optimization loss for both context and prediction stage learning consistently and concurrently. Each HCN stage is trained to optimize the inference of J ground-truth confidence maps. We supervise the model optimization process to maximize the inference discrimination by deploying a separate MSE loss function to the output M_s of each stage s. In particular, we generate the ground-truth confidence map $\tilde{\boldsymbol{m}}_i$ for each single joint i ($i \in \{1, \cdots, J\}$) by placing a Gaussian distribution centered at the labeled location $\boldsymbol{z}_i = (u, v)$. A Gaussian heatmap $\tilde{\boldsymbol{m}}_i$ to represent the i-th joint label is defined as:

$$\tilde{\boldsymbol{m}}_i(j,k) = \frac{1}{2\pi\sigma^2} \exp\left(\frac{-[(j-u)^2 + (k-v)^2]}{2\sigma^2}\right) \tag{5}$$

where (j, k) denotes the spatial location and σ is the spatial variance ($\sigma = 1$ pixel in our evaluations). The heatmap's spatial size for each stage is identical to the input image size. We formulate the stage-wise MSE loss function as:

$$g_{\text{loss}}(M_s, Z, s) = \frac{1}{J}\sum_{i=1}^{J} \|\boldsymbol{m}_i - \tilde{\boldsymbol{m}}_i\|_2^2 \tag{6}$$

The overall HCN loss function is then formulated as:

$$\mathcal{L}(M, Z) = \frac{1}{S}\sum_{s=1}^{S} g_{\text{loss}}(M_s, Z, s) \tag{7}$$

where $M = \{M_s\}_{s=1}^S$ specifies the predicted confidence maps. We treat all stages equally important in the loss composition for simplicity.

Model Training: The HCN model can be trained using the SGD algorithm in an end-to-end manner. Specifically, three types of gradient are involved in the back-propagation of the HCN loss (Eq. (7)) as below.

(1) The gradient for the s-th stage backbone module parameterized by $\boldsymbol{\theta}_s$ ($s = 1, 2, 3, ..., S$):

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \boldsymbol{\theta}_s} &= \frac{1}{S}\sum_{s'=s}^{S} \frac{\partial \mathcal{L}_{s'}^{\text{ctx}}}{\partial \boldsymbol{\theta}_s} \\ &= \frac{1}{S}\sum_{s'=s}^{S} \frac{\partial \mathcal{L}_{s'}^{\text{ctx}}}{\partial h_{s'}^{\text{ctx}}} \frac{\partial h_{s'}^{\text{ctx}}}{\partial \boldsymbol{\theta}_s}\end{aligned} \tag{8}$$

The updating of s-th stage's parameters relies on the gradient of higher-level stages.

(2) The gradient for the CI module parameterized by $\boldsymbol{\gamma}_s$:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \boldsymbol{\gamma}_s} &= \frac{1}{S}\sum_{s'=s}^{S} \frac{\partial \mathcal{L}_{s'}^{\text{ctx}}}{\partial \boldsymbol{\gamma}_s} \\ &= \frac{1}{S}\sum_{s'=s}^{S} \frac{\partial \mathcal{L}_{s'}^{\text{ctx}}}{\partial h_{s'}^{\text{ctx}}} \frac{\partial h_{s'}^{\text{ctx}}}{\partial g_{s'}^{\text{ctx}}} \frac{\partial g_{s'}^{\text{ctx}}}{\partial \boldsymbol{\gamma}_s}\end{aligned} \tag{9}$$

where $s = 2, 3, ..., S$.

(3) The gradient for the LLF module parameterized by $\boldsymbol{\eta}$:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \boldsymbol{\eta}} &= \frac{1}{S}\sum_{s'=1}^{S} \frac{\partial \mathcal{L}_{s'}^{\text{ctx}}}{\partial \boldsymbol{\eta}} \\ &= \frac{1}{S}\sum_{s'=1}^{S} \frac{\partial \mathcal{L}_{s'}^{\text{ctx}}}{\partial h_{s'}^{\text{ctx}}} \frac{\partial h_{s'}^{\text{ctx}}}{\partial g_{s'}^{\text{ctx}}} \frac{\partial g_{s'}^{\text{ctx}}}{\partial f_{\text{llv}}} \frac{\partial f_{\text{llv}}}{\partial \boldsymbol{\eta}}\end{aligned} \tag{10}$$

which is collectively constrained by the loss of all stages due to the stage sharing nature (Eq. (4)).

Model Deployment: Once the HCN model is trained, we can deploy it for pose estimation to a given test image. As in training, we first scale the test image into S different resolutions to match the resolution. Then, we feed them into the corresponding stages and obtain the predicted joint confidence maps at the last prediction stage.

Remarks: HCN is designed specially for efficient pose inference by hierarchical context learning. This not only enables rapid model training and test with lower model inference cost, but also retains the model generalization capability as compared to the common *stacking* scheme. Moreover, HCN is generic in terms of backbone module design. In the modeling idea, our method fundamentally differs from [9] and [43] that iteratively refine the prediction errors, *versus* progressively propagating coarse-to-fine context knowledge over HCN stages. HCN also differs to recurrent pose models [4] that consider neither context nor model efficiency.

TABLE 1. Configuration of an HCN model using Hourglass as the backbone module. A total of 8 HG blocks are used. Stage input is in square shape.

Stage	1st Context	2nd Context	3rd Context	4th Context	5th Context	Prediction
Input Size	32	64	96	128	192	256
Low-Level Feature Module	Conv (3x3 stride 2) Conv (3x3 stride 1) Conv (3x3 stride 1) Residual Max-Pooling Residual Residual					
Scales in Hourglass	1, 1	2, 2	2	3	3	4

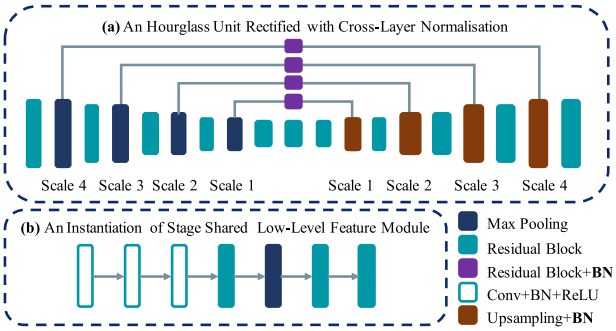


FIGURE 3. HCN Instantiation. (a) Hourglass with 4 scales of representation learning, rectified by adding cross-layer batch normalization to the skip Residual Block and the Upsampling operation. (b) The low-level feature module design.

B. HCN INSTANTIATION

We instantiate a HCN model. The detailed configuration is given in Table 1. This HCN model contains 5 context stages with square input sizes 32/64/96/128/192, and 1 prediction stage with input size 256.

1) LOW-LEVEL FEATURE MODULE

The LLF module (Fig 3(b)) starts with a 3×3 conv layer with stride 2. We use totally three such layers. This both reduces the number of parameters and introduces more non-linearity. We further use three Residual units [17] to gain sufficient learning capacity.

2) BACKBONE MODULE

For the backbone modules (the blue boxes in Fig 2), we select the Hourglass [26] which has proven effective in state-of-the-art models [12], [13], [35]. The Hourglass itself involves a loss term which can be directly integrated into the HCN loss (Eq. (7)).

3) EFFICIENT HIERARCHICAL CONTEXT

HCN aims for efficiency enhancement. The original Hourglass is hence too heavy due to 4 scales of representation learning (Fig 3(a)). To alleviate this issue, we reduce the scale depth of Hourglass for all context stages. Given the coarse-to-fine structure of HCN, we allocate fewer scales to higher-level context stages (Table 1). This respects the ''preattentive'' concept in human visual perception that more efficient and simpler features are processed in the starting stage [24].

4) HOURGLASS IMPROVEMENT

We improve Hourglass's feature fusion. In particular, for each scale, we additionally add batch normalization on top-down and bottom-up features before combining them by element-wise addition (Fig 3). Note, this improvement is not for model efficiency.

This scheme eliminates the cross-layer feature discrepancy whilst strengthens the model training stability. We verify the effect of this refinement in experiments (Table 7). We further introduce a cross-layer batch normalization to mitigate the feature discrepancy issue between layers (Fig 3(a)).

Discussion: Hourglass involves multi-scale feature learning, appearing very similar to the proposed multi-resolution context input in HCN stages. But, they are *conceptually* different. In particular, learning one scale in Hourglass completely relies on the previous scale. which is not context constrained recursive learning as the HCN performs. In contrast to HCN learns from multi-resolution contextual images, Hourglass conducts a multi-scale learning only on the single-resolution observation. Meanwhile, it is due to these intrinsic differences that make HCN and Hourglass complementary to each other and able to contribute respectively in a unified model as shown in our evaluations.

IV. EXPERIMENTS

A. EXPERIMENT SETUP

1) DATASETS

For model evaluation, we used two challenging human pose datasets: MPII [1] and Leeds Sports Pose (LSP) [22]. The MPII includes 40,522 persons in 24,920 images with arbitrary occlusion and background clutters, inter-person interaction, various clothing outfits, and unknown scale variation. MPII images present a wide variety of daily activities in scenes and therefore high challenges for pose estimation. We adopted the standard 25,863/2,958/11,701 train/valid/test split [39]. The LSP dataset consists of 12,000 images captured in different sport events thus presenting rare and challenging poses.

2) EVALUATION METRICS

For MPII and LSP, we used the common pose accuracy performance metrics, i.e. the Percentage of Correct Key-

TABLE 2. Evaluation on the test set of MPII. ***Performance Metric***: PCKh@0.5 and AUC. ***Train/Test Cost Metric***: FLOPs. M = 10^6; G = 10^9; P = 10^{15}. "-": Not reported or lacking details to obtain.

Method	Head	Sho.	Elbo.	Wri.	Hip	Knee	Ank.	*Mean*	*AUC*	*Param*	*Train Cost*	*Test Cost*
Hu et al., [18] CVPR'16	95.0	91.6	83.0	76.6	81.9	74.5	69.5	82.4	51.1	-	-	-
Pishchulin et al., [31] CVPR'16	94.1	90.2	83.4	77.3	82.6	75.7	68.6	82.4	56.5	-	-	-
Lifshitz et al., [25] ECCV'16	97.8	93.3	85.7	80.4	85.3	76.6	70.2	85.0	56.8	76M	-	503G
Gkioxary et al., [16] ECCV'16	96.2	93.1	86.7	82.1	85.2	81.4	74.1	86.1	57.3	-	-	-
Rafi et al., [32] BMVC'16	97.2	93.9	86.4	81.3	86.8	80.6	73.4	86.3	57.3	56M	87P	28G
Belagiannis&Zisserman, [4] FG'17	97.7	95.0	88.2	83.0	87.9	82.6	78.4	88.1	58.8	17M	124P	95G
Insafutdinov et al., [20] ECCV'16	96.8	95.2	89.3	84.4	88.4	83.4	78.0	88.5	60.8	66M	286P	286G
Wei et al., [43] CVPR'16	97.8	95.0	88.7	84.0	88.4	82.8	79.4	88.5	61.4	31M	13,478P	351G
Bulat&Tzimiropoulos, [6] ECCV'16	97.9	95.1	89.9	85.3	89.4	85.7	81.7	89.7	59.6	76M	242P	67G
Newell et al., [26] ECCV'16	98.2	96.3	91.2	87.1	90.1	87.4	83.6	90.9	62.9	26M	343P	55G
Ning et al., [28] TMM'17	98.1	96.3	92.2	87.8	90.6	87.6	82.7	91.2	63.6	74M	964P	124G
Chu et al., [13] CVPR'17	98.5	96.3	91.9	88.1	90.6	88.0	85.0	91.5	63.8	58M	399P	128G
Peng et al., [29] CVPR'18	98.1	96.6	92.5	88.4	90.7	87.7	83.5	91.5	-	26M	-	55G
Chen et al., [11] ICCV'17	98.1	96.5	92.5	88.5	90.2	**89.6**	86.0	91.9	61.6	-	-	-
Yang et al., [44] ICCV'17	98.5	96.7	92.5	88.7	91.1	88.6	86.0	92.0	64.2	28M	238P	46G
Ke et al., [23] ECCV'18	98.5	96.8	92.7	88.4	90.6	89.3	**86.3**	92.1	63.8	-	-	-
Tang et al., [36] ECCV'18	98.4	96.9	92.6	88.7	**91.8**	89.4	86.2	92.3	64.8	16M	219P	34G
Nie et al., [27] CVPR'18	**98.6**	**96.9**	**93.0**	**89.1**	91.7	89.0	86.2	**92.4**	**65.9**	26M	407P	63G
HCN (Ours)	98.3	96.0	90.9	86.7	90.2	87.0	83.6	90.8	63.1	**16M**	**70P**	**18G**

points (PCK) measure that quantifies the percentage of correct detection falling in an error tolerance r [46]. The tolerance r is a normalized quantity w.r.t. the size of torso (PCK@0.2, $r=0.2$ for LSP) or head, denoted as PCKh@0.5. (PCKh@0.5, $r = 0.5$ for MPII). We measured per-joint PCK scores. By varying r, we can obtain a PCK curve and further use the Area Under Curve (AUC) as a more comprehensive metric.

Besides accuracy measurement, we also considered model training and test efficiency. We used the *FLoating point OPerations* (FLOPs) as the metric. Specifically, we measured the model test cost by the FLOPs required to `forward` an image through the model, i.e. `forward-FLOPs`. For training, the cost additionally relies on `batch-size` and `iterations`. We hence measured the training cost as: `batch-size` × `iterations` × `forward-FLOPs`.

3) IMPLEMENTATION DETAILS

All the training and test images are cropped according to the provided positions and scales. Data augmentation includes scaling, rotation, flipping, and color noise addition. We adopted the RMSProp optimizer to train the HCN models. We set the learning rate to 2.5×10^{-4}, the mini-batch size to 4, and the epochs to 150/70 for MPII/LSP.

B. EVALUATION ON MPII

The comparisons between the HCN and 18 state-of-the-art methods on MPII are shown in Table 2. We have two overall observations: **(1)** The HCN model achieves the best training (90 P-FLOPs) and test efficiency (18 G-FLOPs), whilst simultaneously yielding very competitive pose estimation accuracy. **(2)** The HCN is most lightweight with the smallest model size. Consequently, our model not only enables more economical deployments on resource-limited platforms, but also presents less stringent hardware requirements for model training. This significantly improves the cost-effectiveness for large scale deployments in realistic applications.

More specifically, we make three comparisons.

1) Whilst the most accurate model [27] outperforms the HCN by 1.6% (92.4-90.8) in mean PCKh@0.5 and 2.8 (65.9-63.1) in AUC, it is 5.8 (407/70) times slower in training, and 3.5(63/18) times slower in test. Moreover, this model is more complex in design than HCN therefore potentially harder to train.
2) In comparison to the most efficient competitor [32], our model is clearly superior in training (70P *vs* 87P) and test (18G *vs* 28G) costs, in addition to the accuracy performance (90.8 *vs* 86.3 in mean PCKh@0.5, 63.1 *vs* 57.3 in AUC).
3) Compared to Hourglass [26], HCN saves 79.6% ((343-70)/343) training cost whilst attaining 3.1 times (55/18) inference speedup in test. Critically, these advantages are achieved with little model generalization sacrifice.

These evidences indicate that, the proposed HCN provides a good trade-off between accuracy and efficiency in deployment whilst enjoying faster model training advantages.

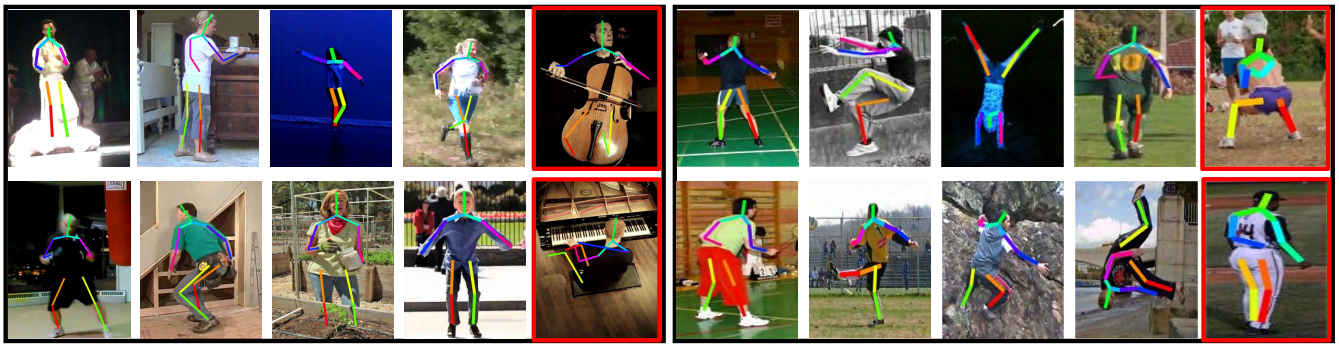
Qualitative Evaluation: To perform visual evaluation, we showed pose estimation results on MPII by our HCN model in Fig 4. It is evident that our model provides accurate pose recognition on images even with highly varying postures against poor illumination, background clutters and occlusions. Meanwhile, the proposed model is likely to fail on extremely challenging cases with missing parts, rare poses, or severe occlusions.

C. EVALUATION ON LSP

We compared the HCN with the state-of-the-art methods on the LSP benchmark in Table 3. In the standard evaluation setting (the top part), it is evident that our method is considerably superior in training efficiency, e.g. accelerating the model training by 62.2% ((37-14)/37) compared to the

TABLE 3. Evaluation on the test set of LSP. *Performance Metric*: PCK@0.2 and AUC. *Train/Test Cost Metric*: FLOPs. M = 10^6; G = 10^9; P = 10^{15}. "-": Not reported or lacking details to obtain. "*": Additionally using the MPII training set.

Method	Head	Sho.	Elbo.	Wri.	Hip	Knee	Ank.	*Mean*	*AUC*	*Param*	*Train Cost*	*Test Cost*
Wang&Li, [42] CVPR'13	84.7	57.1	43.7	36.7	56.7	52.4	50.8	54.6	31.3	-	-	-
Pishchulin et al., [30] ICCV'13	87.2	56.7	46.7	38.0	61.0	57.5	52.7	57.1	35.8	-	-	-
Tompson et al., [40] NIPS'14	90.6	79.2	67.9	63.4	69.5	71.0	64.2	72.3	47.3	-	-	-
Fan et al., [15] CVPR'15	92.4	75.2	65.3	64.0	75.7	68.3	70.4	73.0	43.2	-	-	-
Carreira et al., [9] CVPR'16	90.5	81.8	65.8	59.8	81.6	70.6	62.0	73.1	41.5	-	-	-
Chen&Yuille, [10] NIPS'14	91.8	78.2	71.8	65.5	73.3	70.2	63.4	73.4	40.1	-	-	-
Yang et al., [45] CVPR'16	90.6	78.1	73.8	68.8	74.8	69.9	58.9	73.6	39.3	-	-	-
Rafi et al., [32] BMVC'16	95.8	86.2	79.3	75.0	86.6	83.8	79.8	83.8	56.9	56M	37P	28G
Yu et al., [47] ECCV'16	87.2	88.2	82.4	76.3	**91.4**	85.8	78.7	84.3	55.2	-	-	-
HCN (Ours)	**97.1**	**89.8**	**83.3**	**79.2**	89.0	**86.5**	**84.0**	**87.0**	**59.5**	**16M**	**14P**	**18G**
Chu et al., [13]* CVPR'17	98.1	93.7	89.3	86.9	93.4	94.0	92.5	92.6	64.9	58M	-	128G
Yang et al., [44]* ICCV'17	**98.3**	**94.5**	**92.2**	88.9	94.4	**95.0**	**93.7**	**93.9**	68.5	28M	-	46G
Ning et al., [28]* TMM'17	98.2	94.4	91.8	**89.3**	**94.7**	**95.0**	93.5	**93.9**	**69.1**	74M	-	124G
HCN* (Ours)	98.2	93.8	88.3	86.1	93.0	92.9	92.0	92.0	65.4	**16M**	**14P**	**18G**



MPII LSP

FIGURE 4. Qualitative evaluation of HCN on MPII and LSP. Failure cases indicated by red bounding box.

TABLE 4. Effect of HCN stage design on the MPII val set. *Performance Metric*: PCKh@0.5 and AUC. M = 10^6; G = 10^9; P = 10^{15}.

Stages	#Hourglass / Stage	#Scale / Hourglass	*Mean*	*AUC*	*Param*	*Train Cost*	*Test Cost*
4	2,2,3,1	2,2,3,4	91.0	63.9	18M	**66P**	**17G**
6	2,2,1,1,1,1	1,2,2,3,3,4	**91.5**	64.3	**16M**	70P	18G
8	All 1	1,2,2,3,3,3,3,4	**91.5**	**64.5**	18M	97P	25G

TABLE 5. Effect of batch normalization (BN) in the Context Incorporation module on the MPII val set. *Performance Metric*: PCKh@0.5 and AUC.

BN	Head	Sho.	Elbo.	Wri.	Hip	Knee	Ank.	*Mean*	*AUC*
✗	**98.2**	**96.8**	**91.8**	**88.2**	**89.8**	**87.4**	**85.5**	**91.5**	**64.3**
✓	97.9	96.3	91.2	87.6	89.7	86.4	84.4	90.9	63.8

efficient alternative method [32]. The HCN also yields the best accuracy, improving the mean PCK@0.2 from 84.3% by [47] to 87.0% (+2.7% gain), and the AUC from 55.2 to 59.5 (+4.3 gain). When using the MPII training set for data augmentation (the bottom part), we obtained similar observations. These suggest the performance advantages of our method in the challenging sport event scenarios in addition to the diverse activity settings presented in the MPII test.

D. FURTHER ANALYSIS AND DISCUSSIONS

We conducted a set of model analysis to give further comparisons and model insight on the MPII benchmark [1].

1) STAGE DESIGN

We evaluated the structure of HCN context stages. In particular, we compare three HCN variants with 4/6/8 stages under the constraint that each model consists of the same number of (8 in this case) Hourglass modules with increasing scales from the first to last stages. Table 4 suggests that the design of 6 stages is a good choice giving the best trade-off between model performance and cost.

2) BATCH NORMALISATION IN CI MODULE

Recall that in designing the Context Incorporation (CI) module, we do not apply the Batch Normalization (BN) in prior to

TABLE 6. Effect of sharing low-level feature on the MPII val set. *Performance Metric*: PCKh@0.5 and AUC.

Share	Head	Sho.	Elbo.	Wri.	Hip	Knee	Ank.	*Mean*	*AUC*
✗	97.8	96.1	91.2	87.8	89.4	87.0	85.0	91.0	64.2
✓	**98.2**	**96.8**	**91.8**	**88.2**	**89.8**	**87.4**	**85.5**	**91.5**	**64.3**

TABLE 7. Effect of cross-layer normalization (CLN) on the MPII val set. *Performance Metric*: PCKh@0.5 and AUC.

CLN	Head	Sho.	Elbo.	Wri.	Hip	Knee	Ank.	*Mean*	*AUC*
✗	97.9	96.1	91.2	87.4	89.2	85.4	83.4	90.5	62.9
✓	**98.2**	**96.8**	**91.8**	**88.2**	**89.8**	**87.4**	**85.5**	**91.5**	**64.3**

fusing the context with low-level visual feature. We evaluated this design by comparing with a HCN variant when BN is applied. Table 5 shows that if using the BN for pre-fusion feature normalization, the model performance notably degrades, e.g. −0.6% (91.5-90.9) in mean PCKh@0.5 and −0.5 (64.3-63.8) in AUC. This validates our consideration that the scale information of low-level visual features and context should be preserved in context incorporation at each stage.

3) SHARING LOW-LEVEL FEATURES

We evaluated the performance effect of sharing low-level feature across all HCN stages. Table 6 shows that this design brings +0.5% (91.5-91.0) gain in mean PCKh@0.5 and +0.1 (64.3-64.2) gain in AUC. Moreover, this reduces the parameter number, leading to a more concise pose model.

4) CROSS-LAYER NORMALIZATION

We evaluated the cross-layer normalization design newly introduced to the Hourglass (Fig 3(a)). Table 7 shows that this normalization is effective in improving pose estimation performance, leading to +1.0% (91.5-90.5) increase in mean PCKh@0.5 and +1.4 (64.3-62.9) increase in AUC. This validates our design motivation that the feature scale discrepancy problem in cross-layer fusion may hinder model optimization therefore yielding less discriminative generalization.

V. CONCLUSION

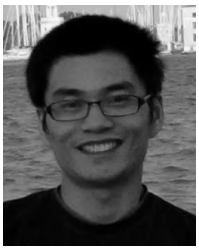
In this work, we have presented a novel and generic Hierarchical Context Network (HCN) for efficiently training and deploying deep human pose estimation models. In principle, this method simulates the coarse-to-fine perception mechanism inherent to the human visual system. This is in contrast to most existing pose methods typically ignoring the crucial model efficiency aspect and focusing only on boosting the accuracy rates. Crucially, HCN is on par with *non-efficient* alternative methods in model generalization capability whilst enjoying significant training and test efficiency advantages. We produced a HCN instantiation model using the Hourglass with extra design improvements. Extensive comparative evaluations have been conducted on three human pose benchmarks to validate the advantages of HCN over a wide range of state-of-the-art methods in challenging daily activity and sporting event scenarios. We performed detailed model component analysis and shed insight into the model performance advantages and design of HCN.

REFERENCES

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. CVPR*, Jun. 2014, pp. 3686–3693.
[2] M. Andriluka, S. Roth, and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in *Proc. CVPR*, Jun. 2009, pp. 1014–1021.
[3] M. Andriluka, S. Roth, and B. Schiele, "Monocular 3D pose estimation and tracking by detection," in *Proc. CVPR*, Jun. 2010, pp. 623–630.
[4] V. Belagiannis and A. Zisserman, "Recurrent human pose estimation," in *Proc. FG*, May 2017, pp. 468–475.
[5] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a 'siamese' time delay neural network," in *Proc. NIPS*, 1994, pp. 737–744.
[6] A. Bulat and G. Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression," in *Proc. ECCV*, 2016, pp. 717–732.
[7] A. Bulat and G. Tzimiropoulos, "Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources," in *Proc. ICCV*, Oct. 2017, pp. 3706–3714.
[8] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. CVPR*, Jul. 2017, pp. 7291–7299.
[9] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, "Human pose estimation with iterative error feedback," in *Proc. CVPR*, Jun. 2016, pp. 4733–4742.
[10] X. Chen and A. L. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," in *Proc. NIPS*, 2014, pp. 1736–1744.
[11] Y. Chen, C. Shen, X.-S. Wei, L. Liu, and J. Yang, "Adversarial posenet: A structure-aware convolutional network for human pose estimation," in *Proc. ICCV*, Oct. 2017, pp. 1212–1221.
[12] X. Chu, W. Ouyang, H. Li, and X. Wang, "Structured feature learning for pose estimation," in *Proc. CVPR*, Jun. 2016, pp. 4715–4723.
[13] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proc. CVPR*, Jul. 2017, pp. 1831–1840.
[14] T. Evgeniou and M. Pontil, "Regularized multi–task learning," in *Proc. SIGKDD*, 2004, pp. 109–117.
[15] 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 *Proc. CVPR*, Jun. 2015, pp. 1347–1355.
[16] G. Gkioxari, A. Toshev, and N. Jaitly, "Chained predictions using convolutional neural networks," in *Proc. ECCV*, 2016, pp. 728–743.
[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Jun. 2016, pp. 770–778.
[18] P. Hu *et al.*, "Bottom-up and top-down reasoning with hierarchical rectified gaussians," in *Proc. CVPR*, Jun. 2016, pp. 5600–5609.
[19] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," in *Proc. CVPR*, Jul. 2017, pp. 4700–4708.
[20] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "DeeperCut: A deeper, stronger, and faster multi-person pose estimation model," in *Proc. ECCV*, 2016, pp. 34–50.

[21] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. ICML*, 2015, pp. 1–11.
[22] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *Proc. BMVC*, 2010, pp. 1–5.
[23] L. Ke, M.-C. Chang, H. Qi, and S. Lyu, "Multi-scale structure-aware network for human pose estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 713–728.
[24] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," in *Matters of Intelligence*. Dordrecht, The Netherlands: Springer, 1987, pp. 115–141. [Online]. Available: https://link.springer.com/chapter/10.1007/978-94-009-3833-5_5#citeas
[25] I. Lifshitz, E. Fetaya, and S. Ullman, "Human pose estimation using deep consensus voting," in *Proc. ECCV*, 2016, pp. 246–260.
[26] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. ECCV*, 2016, pp. 483–499.
[27] X. Nie, J. Feng, Y. Zuo, and S. Yan, "Human pose estimation with parsing induced learner," in *Proc. CVPR*, Jun. 2018, pp. 2100–2108.
[28] G. Ning, Z. Zhang, and Z. He, "Knowledge-guided deep fractal neural networks for human pose estimation," *IEEE Trans. Multimedia*, vol. 20, no. 5, pp. 1246–1259, May 2018.
[29] X. Peng, Z. Tang, F. Yang, R. S. Feris, and D. Metaxas, "Jointly optimize data augmentation and network training: Adversarial data augmentation in human pose estimation," in *Proc. CVPR*, Jun. 2018, pp. 2226–2234.
[30] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele, "Poselet conditioned pictorial structures," in *Proc. CVPR*, Jun. 2013, pp. 588–595.
[31] L. Pishchulin *et al.*, "DeepCut: Joint subset partition and labeling for multi person pose estimation," in *Proc. CVPR*, Jun. 2016, pp. 4929–4937.
[32] U. Rafi, B. Leibe, J. Gall, and I. Kostrikov, "An efficient convolutional network for human pose estimation," in *Proc. BMVC*, 2016, pp. 1–11.
[33] R. P. Rao, G. J. Zelinsky, M. M. Hayhoe, and D. H. Ballard, "Eye movements in visual cognition: A computational study," *Urbana*, vol. 51, p. 61801, Mar. 1997.
[34] L. Sigal and M. J. Black, "Measure locally, reason globally: Occlusion-sensitive articulated pose estimation," in *Proc. CVPR*, Jun. 2006, pp. 2041–2048.
[35] K. Sun, C. Lan, J. Xing, W. Zeng, D. Liu, and J. Wang, "Human pose estimation using global and local normalization," in *Proc. ICCV*, Oct. 2017, pp. 5599–5607.
[36] W. Tang, P. Yu, and Y. Wu, "Deeply learned compositional models for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 197–214.
[37] Z. Tang, X. Peng, S. Geng, L. Wu, S. Zhang, and D. Metaxas, "Quantized densely connected u-nets for efficient landmark localization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 339–354.
[38] Z. Tang, X. Peng, S. Geng, Y. Zhu, and D. N. Metaxas, "Cu-net: Coupled u-nets," in *Proc. BMVC*, 2018, pp. 1–11.
[39] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *Proc. CVPR*, Jun. 2015, pp. 648–656.
[40] 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 *Proc. NIPS*, 2014, pp. 1799–1807.
[41] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proc. CVPR*, Jun. 2014, pp. 1653–1660.
[42] F. Wang and Y. Li, "Beyond physical connections: Tree models in human pose estimation," in *Proc. CVPR*, Jun. 2013, pp. 596–603.
[43] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. CVPR*, Jun. 2016, pp. 4724–4732.
[44] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," in *Proc. ICCV*, Oct. 2017, pp. 1281–1290.
[45] 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 *Proc. CVPR*, Jun. 2016, pp. 3073–3082.
[46] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures of parts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2878–2890, Dec. 2013.
[47] X. Yu, F. Zhou, and M. Chandraker, "Deep deformation network for object landmark localization," in *Proc. ECCV*, 2016, pp. 52–70.

FENG ZHANG received the B.S. degree in network engineering from the Changshu Institute of Technology, Changshu, China, in 2012. He is currently pursuing the Ph.D. degree with the University of Electronic Science and Technology of China, Chengdu, China. His current research interests include pattern recognition and machine learning.



XIATIAN ZHU received the Ph.D. degree from the Queen Mary University of London. He is currently a Computer Vision Researcher with Vision Semantics Limited, London, U.K. His research interests include computer vision and machine learning. He was a recipient of The Sullivan Doctoral Thesis Prize 2016, and an annual award representing the Best Doctoral Thesis submitted to a U.K. University in computer vision.

MAO YE received the B.S. degree from Sichuan Normal University, Chengdu, China, in 1995, the M.S. degree from the University of Electronic Science and Technology of China, Chengdu, in 1998, and the Ph.D. degree from The Chinese University of Hong Kong, Hong Kong, in 2002, all in mathematics. He has been a short-time Visiting Scholar with the University of Queensland and the University of Pennsylvania. He is currently a Professor and the Director of the CVLab, University of Electronic Science and Technology of China. His research interests include machine learning and computer vision. In these areas, he has published over 90 papers in leading international journals or conference proceedings. He has served on the Editorial Board of *Engineering Applications of Artificial Intelligence*. He was a co-recipient of the Best Student Paper Award at the IEEE ICME 2017.