Hybrid Refinement-Correction Heatmaps for Human Pose Estimation

Aouaidjia Kamel, Bin Sheng, Ping Li, *Member, IEEE*, Jinman Kim, and David Dagan Feng, *Fellow, IEEE*

***Abstract*—In this paper, we present a method (Hybrid-Pose) to improve human pose estimation in images. We adopt Stacked Hourglass Networks to design two convolutional neural network models, RNet for pose refinement and CNet for pose correction. The CNet (Correction Network) guides the pose refinement RNet (Refinement Network) to correct the joint location before generating the final pose. Each of the two models is composed of four hourglasses, and each hourglass generates a group of detection heatmaps for the joints. The RNet model hourglasses have the same structure. However, the CNet model is designed with hourglasses of different structures for pose guidance. Since the pose estimation in RGB images is very sensitive to the image scene, our proposed approach generates multiple outputs of detection heatmaps to broaden the searching scope for the correct joints locations. We use the RNet model to refine the joints locations in each hourglass stage horizontally, then the heatmaps of each stage are fused with the heatmaps of all the CNet model hourglasses vertically in a hybrid manner. Our method shows competitive results with the existing state-of-the-art approaches on MPII and FLIC benchmark datasets. Although our proposed method focuses on improving single-person pose estimation, we also show the influence of this improvement on multi-person pose estimation by detecting multiple people using SSD detector, then estimating the pose of each person individually.**

***Index Terms*—Human pose estimation, pose refinement, pose correction, heatmaps fusion.**

I. Introduction

HUMAN pose estimation is a key to understand human behavior in images. It can be defined as the task of recovering human body parts in the image scene, which is essential for various computer vision applications that require information about people's actions, such as surveillance, human-computer interaction, and robotics. However, estimating body parts in colored images is one of the complexes problems in computer vision for a long time due to many factors including, variation in the background scene, illumination, clothing color, and occlusion with other people or objects, which makes it challenging to come up with a perfect algorithm that can identify the correct location of the body parts in every scene. Another cause of complexity lies in a large number of possible positions of the human body that leads to unlimited searching space.

A vast number of approaches have been reported to solve the problem. The early work of human pose estimation tackled the difficulty by proposing complicated structured models for prediction. The classical popular model is the pictorial structure [1] motivated by [2], it represents the human body as a structured graphical model that expresses kinematic dependencies between body parts. This first initial version of the model was very limited and can only work in a plain background, which is not the case in most images/videos scenes. However, several improved versions of this model have been proposed to enhance the detection in a complex background with difficult body poses. Generally, pictorial structure methods can be in the form of a tree structure like [3], [4] that model the body parts and the connection between them, or a non-tree structure models which add loop branches to detect the parts in more complicated poses [5]–[7]. Besides graphical models, spatial prediction techniques map directly the input image to the body joints coordinates through sequential models of multiple stages [8].

Although the previously mentioned methods achieved relatively good accuracy, they are still far from covering the body poses in complicated image scenes or difficult body poses. In recent years, deep learning and convolutional neural networks (ConvNets) have made a huge success in image recognition. ConvNet is a powerful technique for feature extraction that can automatically learn discriminative features from training data to solve classification and regression problems. It became popular when the ImageNet model [9] was introduced to classify images of many categories. Later on, several ConvNet models have been built on this model to tackle many computer vision problems. Recently, the research in deep learning and image recognition focuses on how to boost the classification and regression accuracy by proposing new architectures and building blocks of ConvNets to improve the performance. ResNet [10], Rethinking Inception [11] and Inception-v4 [12] are successfully proposed

Manuscript received July 12, 2018; revised May 10, 2020; accepted May 14, 2020. Date of publication June 3, 2020; date of current version April 23, 2021. This work was supported in part by the National Key Research and Development Program of China (2018YFF0300903), and in part by the Science and Technology Commission of Shanghai Municipality under Grants 18410750700, 17411952600, and 16DZ0501100, in part by the National Natural Science Foundation of China under Grants 61872241 and 61572316, and in part by The Hong Kong Polytechnic University under Grants P0030419 and P0030929. The associate editor coordinating the review of this manuscript and approving it for publication was Jingdong Wang. *(Corresponding author: Bin Sheng.)*

Aouaidjia Kamel is with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China, and also with the Space Techniques Center, Algerian Space Agency, Arzew 31200, Algeria (e-mail: kameldz40@gmail.com).

Bin Sheng is with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: shengbin@sjtu.edu.cn).

Ping Li is with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong (e-mail: p.li@polyu.edu.hk).

Jinman Kim and David Dagan Feng are with the Biomedical and Multimedia Information Technology Research Group School of Information Technologies,The University of Sydney, Sydney, NSW 2006, Australia (e-mail: jinman.kim@sydney.edu.au; dagan.feng@sydney.edu.au).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2020.2999181

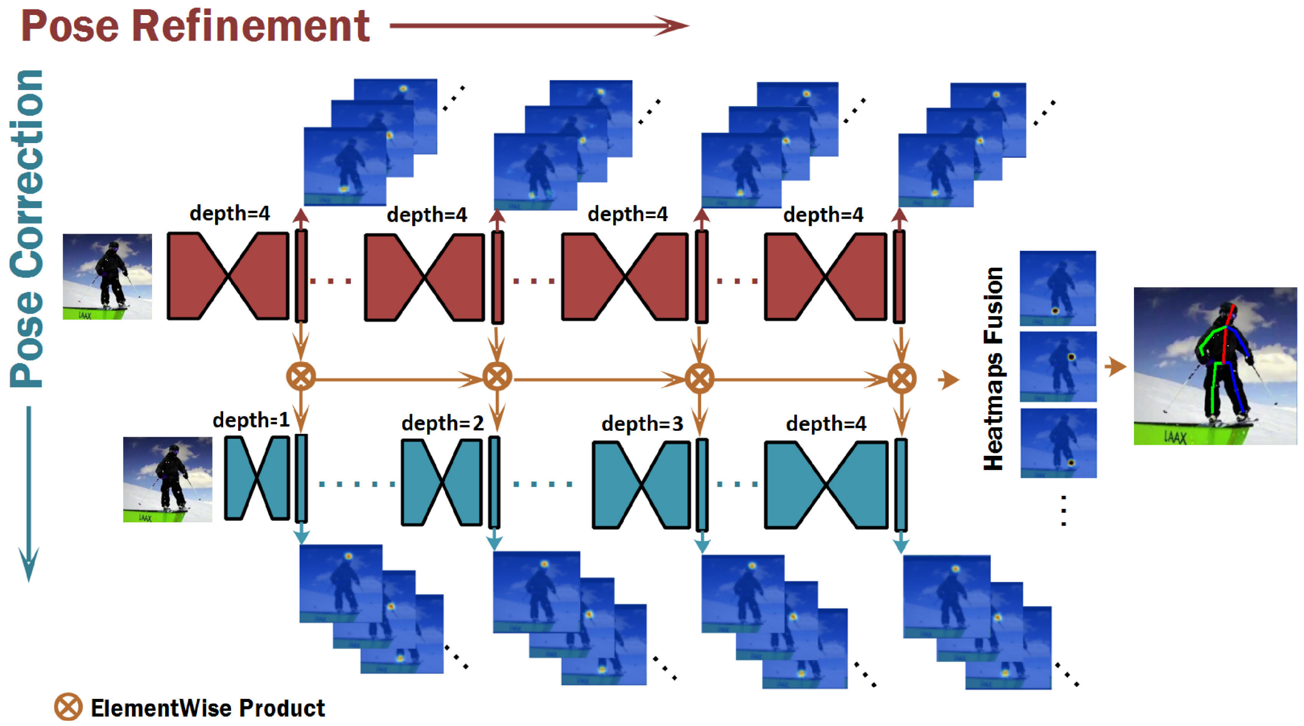


Fig. 1. Our proposed Hybrid Refinement-Correction Heatmaps method for human pose estimation. Two ConvNet models are used for pose refinement and correction. The detection heatmaps are refined horizontally by the hourglasses of the refinement model and corrected vertically by the hourglasses of the correction model to generate final detection heatmaps via a fusion operation.

structures which became later the main components of a new state-of-the-art models. Lately, ConvNets are widely used to understand humans in images and treat complex images/videos recognition problems [13]–[17].

Hourglass networks [18] is one of the recent successful ConvNet structures for human pose estimation that treats the input image in multi-scale by down-sampling the feature maps, then up-sampling them again to capture body features in different scales. Multiple hourglasses are used for pose refinement, and the last hourglass output is expected to generate accurate detection heatmaps of the joints locations. The challenge facing existing methods for human pose estimation is how to deal with difficult situations such as occlusion, different body positions and scales. Although deep learning methods based on ConvNet like [18] and [19] have better performance than the previous classical models, they still fail to predict correct locations of body joints in challenging situations depending on the learned features. The principle of those approaches is treating the input image with multiple processing stages for pose refinement. However, even though the refinement process will not be efficient if the joint is located in the wrong position at an early stage, because the rest of the stages will locate the joint in the coarse initial detected area while the correct location could be in another area of the image. The missing key concept in those approaches is the guidance clue that can help to check the correct location in each stage before moving to the next stage, which leads to consider that the type of features to be learned for guidance should be different from the features learned for refinement. It comes to mind that a simple approach to tackle the problem is by assembling multiple network structures to enrich the feature extraction. However, the assembling has to be based on using the appropriate network structures, optimisation parameters, and learning process to extract the correct pose features. The previous analysis motivated us to think that a better approach should be based on using two ConvNet models, one for pose refinement and another for pose correction.

In this paper, we propose a hybrid pose estimation approach using two ConvNet models, RNet (Refinement Network) for pose refinement and CNet (Correction Network) for pose correction. We exploit the multi-scale processing of the hourglass structure presented in [18] to construct the models. The RNet is designed using four hourglasses of the same structure for pose refinement, each hourglass generates heatmaps of body poses for the next hourglass, and the last hourglass generates the final joints locations. However, the pose correction network CNet is designed with four hourglasses which have different structures (or depths) to capture missing features in the RNet model. In the end, the heatmaps generated from each hourglass of the RNet model are fused with all the heatmaps of the CNet model for pose correction to handle missing features. Fig. 1 presents the idea of the proposed approach.

In the results section, we show that the proposed approach generates competitive pose estimation results on two benchmarks datasets: MPII and FLIC. We also show the influence of the accuracy on the pose estimation in the presence and the absence of the pose correction network. The visual comparison between our two fused models results and [18] shows that in difficult situations, our model performs better. Despite that our main focus in this work is to improve single-person pose estimation, we also show visual results of the influence of this improvement on multi-person pose estimation by detecting multiple people using SSD (Single Shot MultiBox Detector) pre-trained model [20], then we estimate the pose of each person individually using our proposed approach. The main contributions of our work can be summarized as follows:

- **A hybrid approach** that works in two sides using two ConvNet models, $RNet$ for pose refinement and $CNet$ for pose correction. The pose estimation heatmaps are refined along $RNet$ model and the joint location is corrected with the $CNet$ model.
- **A pose correction network (CNet)** is constructed with hourglass structures of different depths to capture another type of features different from those of the pose refinement model. The CNet model helps to compensate the missed features and hence, correct the joint location in case of a wrong estimation by RNet model.

The rest of this paper is organized as follows: Section II discusses the related work for pose estimation. The technical details of the method are given in Section III. Experiments and results are shown in Section IV. Finally, a conclusion summarizes the proposed work in Section V.

II. Related Work

This section reviews some of the recent proposed approaches for pose estimation using deep convolutional neural networks. Carreira *et al.* [21] proposed a framework that exploits a ConvNet for feature extraction to predict the output body poses with a feedback error that is progressively fed again to the input to change the initial solution in an iterative process called Iterative Error Feedback. In [22], deep ConvNets are used with a graphical model to learn the presence of parts and the relationship between them using image batches. Ouyang *et al.* [23] proposed a single deep model for both tasks of human detection and pose estimation using multiple sources of knowledge. Temporal information using multiple video frames is exploited in [24], where they proposed a ConvNet architecture that learns body poses using optical flow images. In [25], two ConvNet models are jointly-trained. The first model generates a coarse body part localisation heatmaps that describe the likelihood of a joint to be in a specific spatial location, it takes three levels of an RGB Gaussian pyramid inputs, each level is processed with a sub-sequence of convolutional and pooling layers, and the output of the three ConvNet branches are fused for further processing with another sequence of layers to generates a heatmap for each joint per-pixel location. The second model refines the joint location by reusing features of convolutional layers from the first model to improve the joints localisation accuracy. DeepPose [26] proposed an efficient and simple deep neural network model that predicts body joints directly from the input image. [27] employs two ConvNet models, one model called Independent Losses Pose Nets (ILPNs) to detect the joints location on a global level, and another model called Convolutional Local Detectors (CLDs) to locate the joints in a potential region.

Convolutional Pose Machines [19] (which is an improved version of [8]) is one of the recent successful pose estimation methods that treat the problem as an iterative process of many stages using a ConvNet model that refines the joints locations at each stage. Based on the same idea of iterative processing and joints locations refinement, in [18], a new ConvNet architecture involves several processing stages is proposed. In contrast to [19] that utilizes the same model in several iterations, [18] used a single model with multiple processing stages of consecutive structures for refinement called hourglasses. The hourglass treats the input image in multi-scales by down-sampling the feature maps, then up-sampling them for the processing again by the next hourglass. In [28], an hourglass-based structure called multi-scale supervision network was proposed. This network is followed by another regression network that combines multi-scale features of the hourglass to improve joints locations. Moreover, a structure-aware loss is set between the hourglasses to ensure learning human skeleton structure.

III. Hybrid Refinement-Correction Method

Our proposed method for pose estimation is a hybrid approach which operates in two directions using two ConvNet models, RNet is used for pose refinement and CNet is employed for pose correction. The final pose estimation result is the fusion of the heatmaps generated from the two models. The RNet and CNet models have different architectures. The two models accept an input image scaled to a size of 256×256, and the main difference between the two networks lies in the hourglasses depth size. We define the hourglass depth d as the number of times the feature maps down-sampled and up-sampled. Fig. 4(a) and Fig. 4(b) shows hourglasses of depth $d = 4$ and $d = 2$ respectively. Both models are composed of four hourglasses, but all the hourglasses of the RNet model has depth $d = 4$, whilst in the CNet model, we use hourglasses of different depths (from $d = 1$ to $d = 4$). This design allows the RNet model to have an architecture of deep hourglasses that refine the joint location four times, which is useful for updating the joints locations. The design of the CNet model allows extracting features from several scales with hourglasses of a different structure from RNet. It provides another type of features useful to get extra information about the poses that may have wrongly estimated by the RNet model. Fig. 3 shows an example that illustrates the advantage of the proposed method, and Fig. 2 shows the detailed structure of the two proposed ConvNet models. In this section, we discuss the network architecture of both RNet and CNet, and we also describe how the detection heatmaps are fused.

A. Pose Refinement Network (RNet)

Fig. 2 (RNet) shows the structure of the RNet model in details including the layer's types, and the connections between layers. The main building block of the hourglass structure is a residual module, which is the reason behind its good performance. The type of the residual module that we proposed and used in this model is presented in Fig. 5(a). Based on the work presented in [12], which showed that different types of residual modules can improve the learning performance, we use a new residual module which has an extra deep branch further to the one used in [18] to enhance the refinement process.

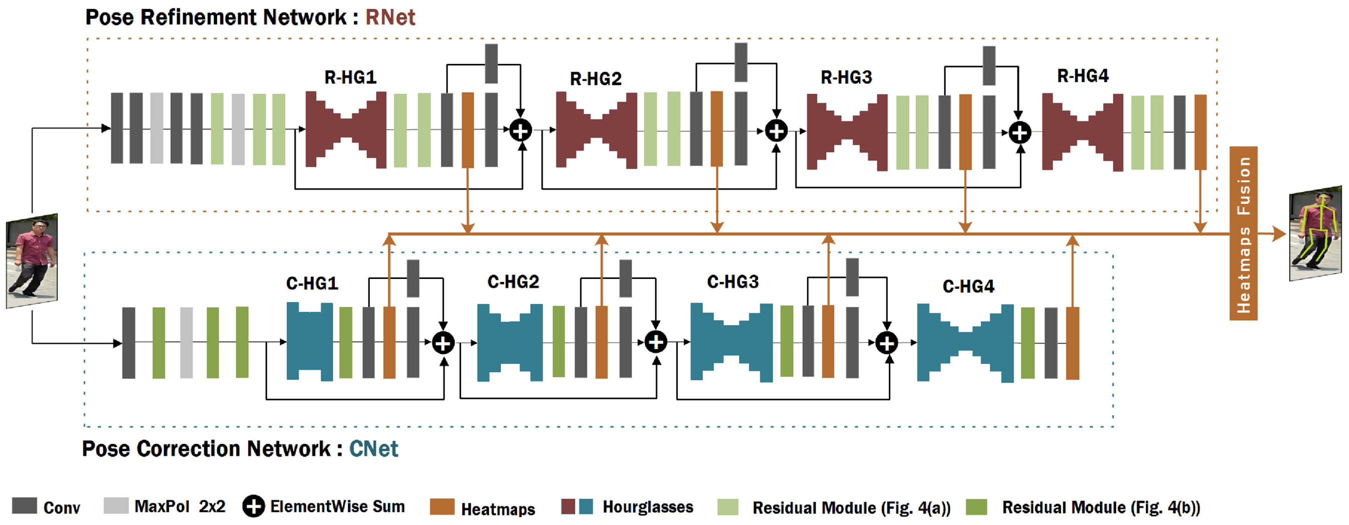

Pose Refinement Network : RNet
R-HG1
R-HG2
R-HG3
R-HG4
Heatmaps Fusion
C-HG1
C-HG2
C-HG3
C-HG4
Pose Correction Network : CNet
Conv
MaxPol 2x2
ElementWise Sum
Heatmaps
Hourglasses
Residual Module (Fig. 4(a))
Residual Module (Fig. 4(b))

Fig. 2. A detailed structure of our proposed two ConvNet models presented in Fig. 1.

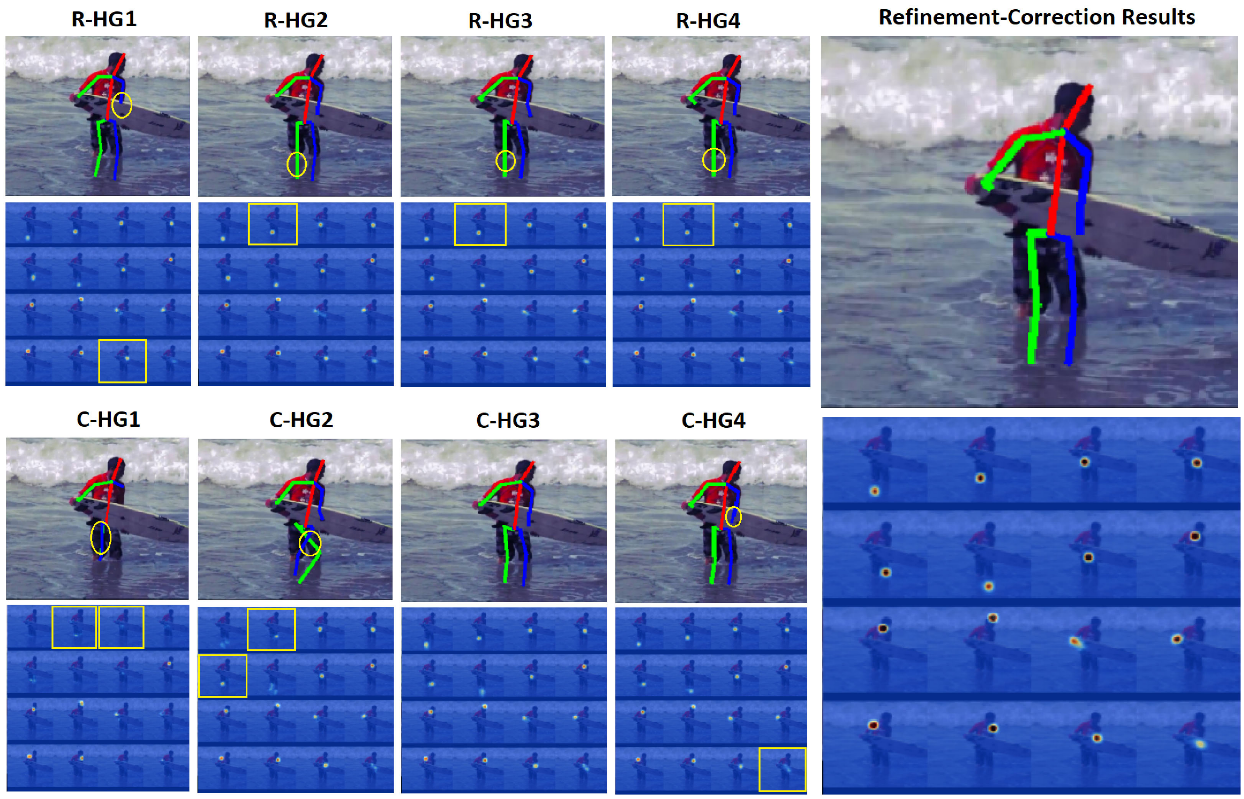


Fig. 3. **Contribution**: An example of heatmaps generated from the different hourglasses presented in Fig. 2 associated with the estimated joints on the original image. The yellow circles indicate the wrong locations of the joints on the image, and the yellow squares show the related heatmaps. The Refinement-Correction result shows accurate pose estimation.

The first hourglass is preceded by nine layers as follows: Two 1×1 convolutional layers followed by a max pooling layer and two other 1×1 convolutional layers. After that, a residual module is used and followed by a max pooling layer, and followed again by two other residual modules. The outputs of this sequence of layers are feature maps of size $64 \times 64 \times 256$, where 64 represents the height and the width of the feature maps, and 256 represents the number of feature maps. These feature maps are used as an input to be processed by the hourglass. The structure of the hourglass used in this model is presented in Fig. 4(a). The residual module (ResNet [10]) helps the network to keep learning the desired features despite a large number of layers due to the shortcut connections. Based on the same idea, we doubled the number of residual modules at each location in the hourglass to two residuals instead of one to make the network deeper for feature extraction without falling into the problem of

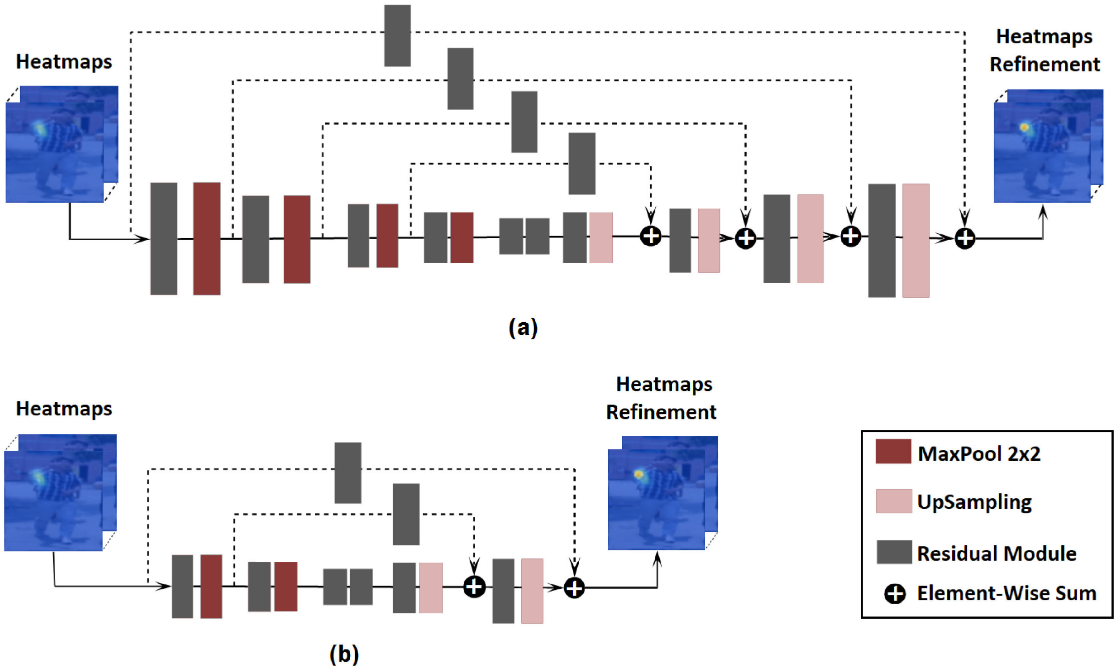


Fig. 4. A detailed hourglass network structure with one residual module. (a) Hourglass of depth d = 4. (b) Hourglass of depth d = 2. The depth refers to how many times that the feature maps are down-sampled and up-sampled.

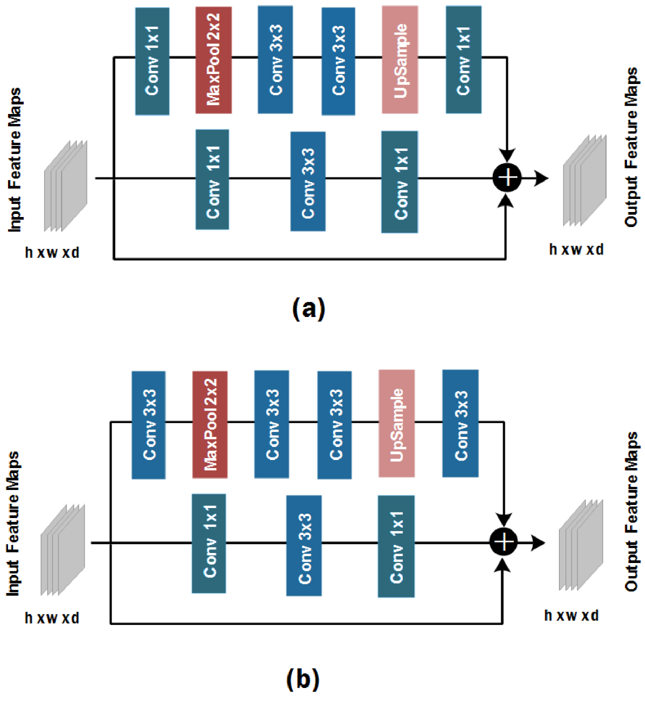


Fig. 5. The two proposed residual modules. (a) The residual module used in the RNet model. (b) The residual module used in the CNet model.

vanishing gradient. The output feature maps size of the hourglass is the same as its input size. Each hourglass output is fed again to two consecutive residual modules followed by a convolutional layer. Finally, a convolutional layer generates heatmaps of size 64 × 64 × 16, where each heatmap represents one joint. The next stage of the new hourglass is started by summing the feature maps of the hourglass input and output. The process is repeated three times before getting the fourth detection heatmaps from the last hourglass.

B. *Pose Correction Network (CNet)*

Since our objective is to design a correction model, we propose a ConvNet architecture to extract features differently from the RNet model. The hourglasses used here are of various depths, unlike the RNet model whose hourglasses have the same depth. Furthermore, the structure of the CNet model differs by the number of layers used before the first hourglass as follows: 7×7 convolution layer is followed by a residual module and a max pooling layer, then followed again by two residual modules. To reduce the computation and at the same time to extract different features from the ones of the RNet model, instead of using two residual modules at each location of the hourglass, we use only one residual module which has a different structure from the one used in the RNet model (Fig. 5(b)). Besides the previously mentioned differences, the rest of the model structure is the same as the RNet model. The Network architecture of the CNet is presented in Fig. 2 (CNet).

C. *Implementation Details*

1) Training Parameters: Instead of training the models jointly, we trained each model separately to make the learnt body features independent because, in case of jointly training, there is only one loss function for both models where the heatmaps prediction results are adjusted in one way. However, with two loss functions, the heatmaps are adjusted differently and offer more variation in the joints locations. The RNet model is trained for 200 epochs with 2000 iterations in each epoch, and with a learning rate of 2.5×10^{-6}. However, the CNet model is trained for 100 epochs with 4000 iterations, and with a learning rate of 2.5×10^{-4}. At first, the learning rate of the RNet model is initialised based on [18] (2.5×10^{-4}) then we found that the loss function behavior is unstable and doesn't decrease smoothly, so we decreased it to 2.5×10^{-6}. But for the CNet model, we kept the same learning rate value which was convenient for a smooth learning process. The number of iterations was chosen to be enough for getting the lowest loss function value. Each model receives a batch size of four images in each iteration. The training has been performed on a GPU of 12,2 GB memory of a machine of Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20 GHz (10 cores), 64 GB of RAM, and 64 bits operating system.

2) Hourglass Calculation: Eq. (1) to Eq. (5) formulate the calculation of an hourglass of depth d for the models RNet and CNet. Eq. (6) and Eq. (7) calculate the heatmaps from the hourglasses of the RNet and the CNet models respectively.

$$Down = Res(Pool(nf, nf), nf) \tag{1}$$

$$Up = Res(UpSample(nf, nf)) \tag{2}$$

$$HG(Res, d, nf, nf) = Up^d(Res^d(Down^d(nf, nf)))$$
$$d = \{1, ..., 4\} \tag{3}$$

$$R - HG_i = HG(Res_{RNet}, d, nf, nf)$$
$$d = 4, \quad i = \{1, ..., 4\} \tag{4}$$

$$C - HG_i = HG(Res_{CNet}, d, nf, nf)$$
$$d = \{1, ..., 4\}, \quad i = \{1, ..., 4\} \tag{5}$$

where *R-HG*$_i$ is the hourglass number i of the RNet model, and *C-HG*$_i$ is the hourglass number i of the CNet model. nf is the number of feature maps of the hourglass layers where $nf = 256$ (the input and the output of the feature maps of the hourglasses have the same number nf). d is the depth of the hourglass. For example, Fig. 4(a) is an hourglass of depth $d = 4$ and Fig. 4(b) is an hourglass of depth $d = 2$. $Down$ decreases the dimension of the feature maps in the hourglass using max pooling operation $Pool$, and Up increases the dimension size through the upsampling operation $UpSample$. Res represents the residual module in general. Res_{RNet} and Res_{CNet} are the residual modules used in the RNet and CNet models respectively, where their structure is presented in Fig. 5. The notation X^d means applying the operation X d times.

$$R - H_i = Cnv^2_{1\times1}(Res^2_{RNet}(R - HG_i, nf), nf) \tag{6}$$

$$C - H_i = Cnv^2_{1\times1}(Res_{CNet}(C - HG_i, nf), nf) \tag{7}$$

$$|R - H_i| = |C - H_i| = 16 \times 64 \times 64 \tag{8}$$

where *R-H*$_i$ and *C-H*$_i$ are the detection heatmaps of the hourglasses *R-HG*$_i$ and *C-HG*$_i$ of the models RNet and CNet respectively. $|R - H_i|$ and $|C - H_i|$ are the sizes of the detection heatmaps which are identical. 64 is the height and the width of the heatmaps, and 16 is the number of heatmaps (number of body joints).

D. Heatmaps Fusion

From the two models, we get eight groups of detection heatmaps. Since the last hourglass ($d = 4$) in each model is the final stage of the refinement process, it is expected that it should generate the most accurate joints locations, which is the case in most predictions results. However, sometimes the first, second, or the third hourglass with a depth $d = 1, d = 2,$ or $d = 3$ respectively, predicts joint locations that are more accurate than those of the last hourglass (Fig. (3)). Furthermore, an hourglass of the CNet model of depth $d = 1$ can generate joints locations that are more accurate than those of an hourglass (of depth $d = 4$) of the RNet model. From the fact that we cannot know which hourglass predicts accurate results, we propose a fusion operation of the detection heatmaps generated from the hourglasses of the two models in a complementary process. We tried many heatmaps fusion based on element-wise operations, and we found that the best operation that generates good results must be performed between the heatmaps of each hourglass of the RNet model and the four heatmaps generated from the hourglasses of the CNet model using the element-wise multiplication, which consists of multiplying the pixels of the same position in the heatmaps. A pixel of high value in the heatmap indicates that it may represent the joint location (areas of high energy in the heatmaps of Fig. 3). Since the multiplication increases the values of pixels of low energy with pixels of high energy, it supports each heatmap of the RNet by the joint location of all the heatmaps of the CNet to strengthen the possibility of getting the correct pixel position (Eq. (9)). The results of the four multiplication operations are summed together using the element-wise addition to generate the final heatmaps that represent the joints locations (Eq. (10)).

Fig. 3 shows a visual explanation of the positive effect of the refinement model RNet and the correction model CNet with the fusion operation to get an accurate prediction. The detection heatmaps of the pose refinement model are almost the same (Fig. 3(Top)). Even the wrong joint location (right knee) is the same for three hourglasses because the RNet model just refines the joints locations of the previous heatmaps, which is only useful when the joint is predicted at the right location at the beginning. However, the CNet model detection heatmaps generated relatively different joint locations for the same joint due to the diversity in its hourglasses depths. C-HG3 and C-HG4 (Fig. 3) show accurate joints locations compared to those of the RNet model, and the fusion of the refinement network and the correction network detection heatmaps reflects the effectiveness of the proposed method (Fig. 3(Refinement-Correction Results)).

$$Mul_k = R - H_k \odot (C - H_1 \odot \ldots \odot C - H_4)$$
$$k = \{1, ..., 4\} \tag{9}$$

$$Fus(RNet, CNet) = Mul_1 \oplus \ldots \oplus Mul_4 \tag{10}$$

where Mul_k is the multiplication of the heatmaps *R-H*$_k$ of the RNet model and the four heatmaps of the CNet model *C-H*$_1$ to *C-H*$_4$. $Fus(RNet, CNet)$ sums all the results of the multiplications to generate the prediction of the final joints locations. $\odot$ indicates the element-wise multiplication, and $\oplus$ refers to the element-wise addition.

TABLE I
SINGLE-PERSON POSE ESTIMATION COMPARISON RESULTS ON MPII VALIDATION SET (PCKH@0.5)

Method	**Head**	**Shoulder**	**Elbow**	**Wrist**	**Hip**	**Knee**	**Ankle**	**Total**
Tompson et al. [25]	96.6	92.7	83.4	76.7	81.9	73.3	65.4	82.3
Newell et al. [18]	**97.1**	**96.1**	**90.8**	**86.2**	**89.9**	**85.9**	**83.5**	**90.0**
Yang et al. [29]	97.4	96.2	91.1	86.9	90.1	86.0	83.9	90.3
Tang et al. [30]	97.4	96.2	91.0	86.9	90.6	86.8	84.5	90.5
Xiao et al. [31]	97.5	96.1	90.5	85.4	90.1	85.7	82.3	90.1
Sun et al. (single-scale) [32]	97.1	95.9	90.3	86.4	89.1	87.1	83.3	90.3
Sun et al. (multi-scale) [32]	97.7	96.3	90.9	86.7	89.7	87.4	84.1	90.8
Our (Hybrid-Pose)	**97.5**	**96.2**	**90.8**	**86.6**	**89.3**	**87.1**	**83.4**	**90.4**

IV. EXPERIMENTAL RESULTS

We evaluate the proposed method on MPII and FLIC datasets for single-person pose estimation. Both datasets provide images in various scenes with multiple people. However, only one person is selected for single-person pose estimation. Although our main focus in this work is to improve single-person pose estimation, we also applied the proposed method on samples of the datasets to estimate the poses of multiple people in the image for visual evaluation. We use state-of-the-art SSD [20] object detection pre-trained model to detect multiple people, then we apply the proposed method to estimate the body pose of each person separately.

A. Evaluation on MPII Dataset

The state-of-the-art benchmark MPII Human Pose dataset [34] includes around 25k images holding over 40k people associated with body joints annotations. The images are collected from YouTube videos with a diversity of human activities in challenging body poses, including overlapping and occlusion. In recent years most of the state-of-the-art pose estimation approaches have been evaluated their methods on this dataset because it provides pose annotations of a large number of images in a variety of human activities, which motivates researchers to work on deep learning models that require a large amount of data for training. Among multiple people in the image, only one person is selected for training and testing. Besides the body joints annotations provided by the MPII dataset, it provides also the scale and the center that define the rough location of the person in each image sample. The target person is cropped and resized to 256×256 for training and testing. We follow the data augmentation in [18] with rotation (+/$-$ 30 degrees) and scaling (0.75–1.25) for training. For single-person pose evaluation, we use PCK (Percentage of Correct Keypoints) metric defined in [35] which measures the percentage of joints localisation within a normalized distance of the ground truth. We also use PCKh metric defined in [34] to set a threshold of the head segment length for matching the joint location to the ground truth.

Table I shows the PCKh results at a threshold of 0.5 on MPII validation set for single-person pose estimation evaluation. We follow the same metrics of the previous methods such as [18] and [32]. We compared the proposed method with existing works that evaluated the performance on the validation set like us. Our method shows a competitive performance with most of the existing state-of-the-art methods on single-person pose estimation except [32] (multi-scale) because of using multi-scale prediction. However, our method performs better compared to their single-scale results. In our work, we focused on the idea of investigating the use of two models to improve the prediction in difficult body poses with the help of the correction network. This idea can be applied to deeper fused models with more hourglasses to improve the accuracy that is shown in Table I. To test the proposed approach in handling difficult body positions, we made a visual comparison between the results of the fusion operation and the method presented in [18] using their pre-trained model of eight hourglasses. The comparison in Fig. 11 shows that our approach can predict body joints even in occlusion cases.

Fig. 8 shows the PCKh detection curves of a normalized distance in an interval of values between 0 and 0.5 of the refinement model RNet, the correction model CNet, and the fusion operation $Fus(RNet, CNet)$ for each joint. Within the change of the normalized distance, the fusion of the two models detection rate is better than using RNet model or CNet model only, and it is also better than the changing rate of the method proposed by [25]. In Fig. 6, we applied the proposed method on images containing people in different poses selected from the testing and the validation sets of MPII dataset. The proposed method shows that it can predict human poses in challenging situations correctly. In Fig. 7, we present the prediction results generated from each of the eight hourglasses of the two models on MPII validation set.

The performance of our proposed single-person pose estimation method is reflected positively for multi-person pose estimation. We also applied the proposed method on images from MPII validation and testing sets of different situations, near and far positions, occlusion, and with different body poses. The visual results on multiple people are shown in Fig. 9. Additionally, in Fig. 10, we show a visual comparison between the results of the proposed method and the results generated by [18] on multiple people in the image, where our method performs better in occlusion and challenging body poses. Fig. 13 (four right images) shows some failure cases of our method on MPII where the body parts are occluded or the person in a very near position.

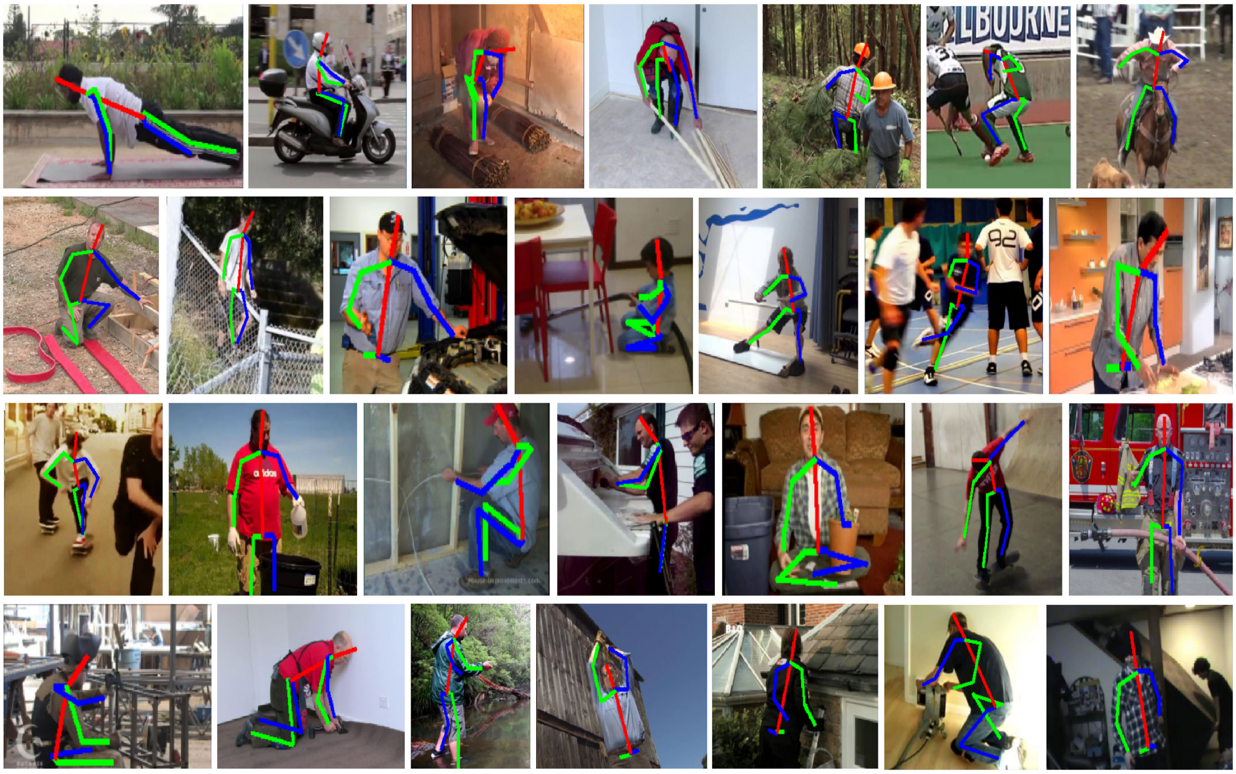


Fig. 6. Single-person pose estimation results on samples from MPII validation and test sets.

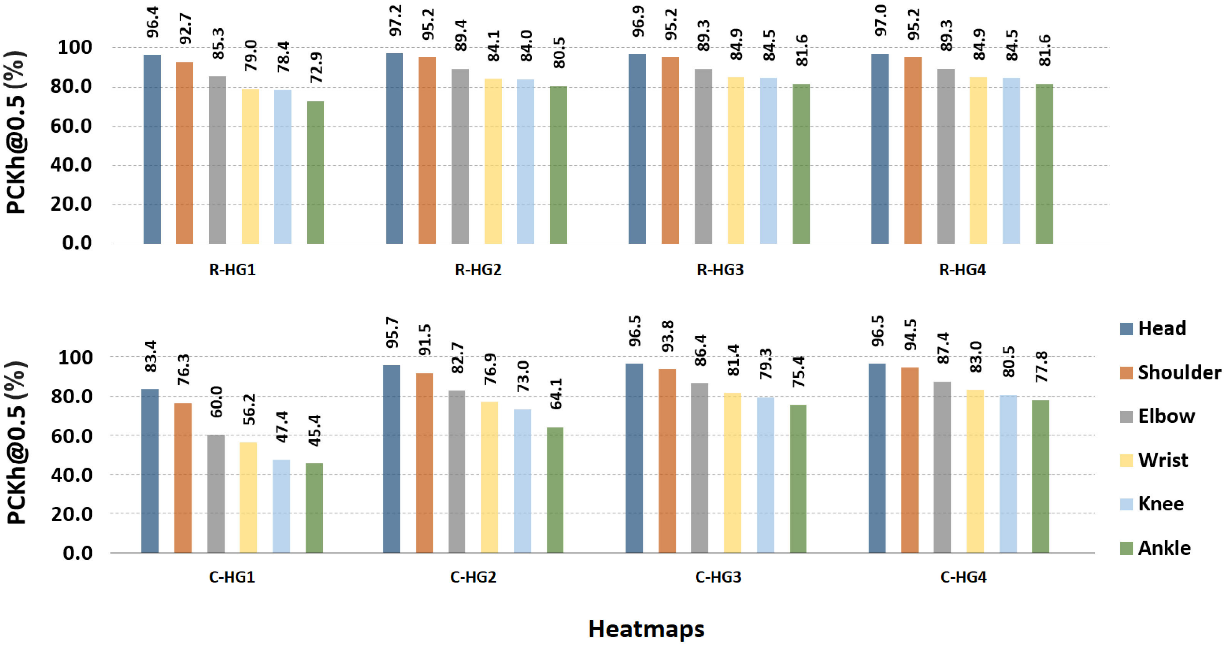


Fig. 7. Prediction accuracy generated from each hourglass of the RNet and CNet models on MPII validation set.

B. Evaluation on FLIC Dataset

FLIC dataset [33] is collected from movies and annotated on the upper body parts. It consists of 5003 images, 3987 images for training and 1016 images for testing. Even though MPII dataset includes images of the upper body, we evaluate the proposed Refinement-Correction approach on the FLIC dataset to test its performance on the upper body where in most cases people are in near positions. We used our trained models on MPII dataset to evaluate the accuracy on the testing set directly without training on FLIC since the MPII dataset includes images of the upper body as well.

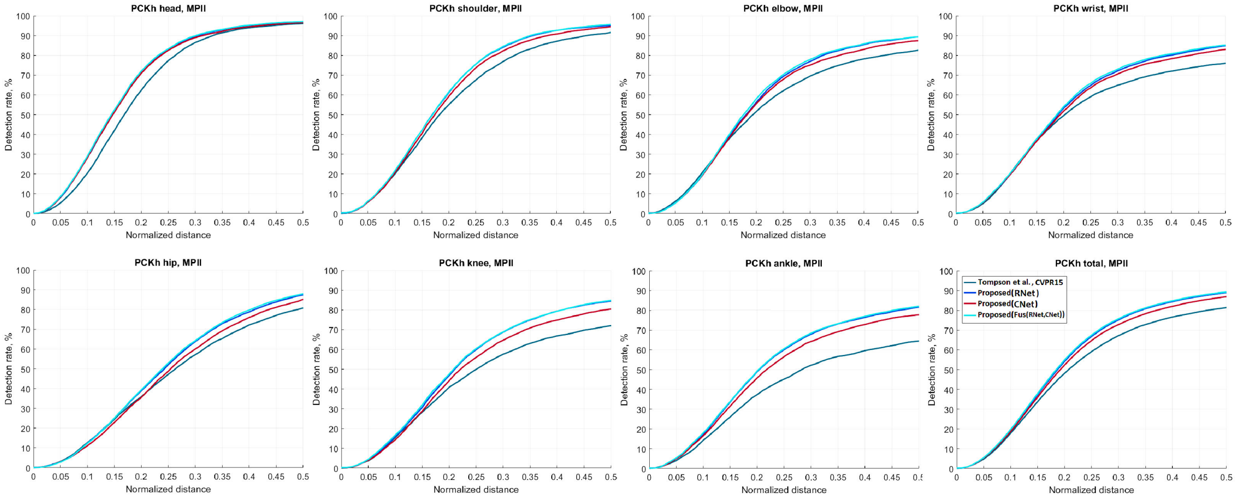


Fig. 8. Single-person PCKh comparison on MPII validation set.

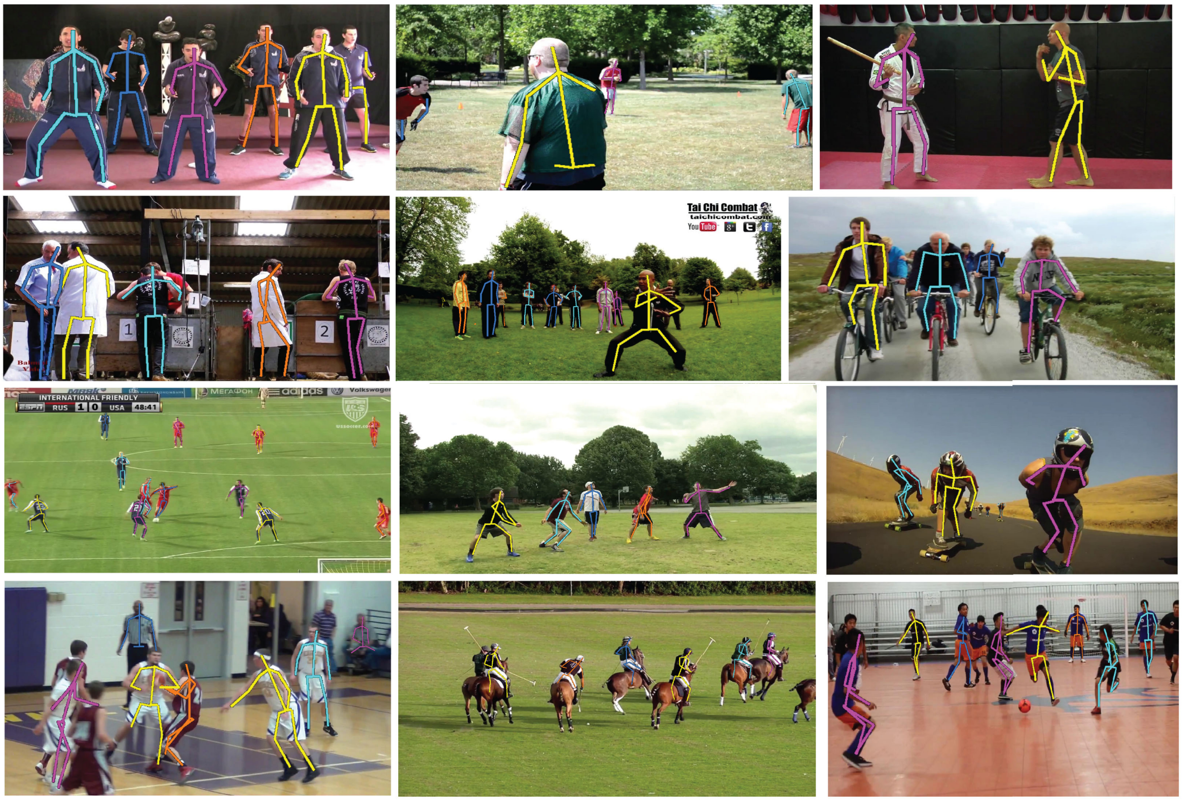


Fig. 9. Multi-Person pose estimation results on samples from MPII validation and testing sets.

Table II shows the comparison results with some of the existing methods on the elbow and wrist joints on single-person pose estimation. The predictions of our method show competitive results on this dataset despite that we did not train the model on FLIC like the compared methods, which is the reason behind the lower accuracy than [18]. We also applied the proposed method to multiple people in images from FLIC testing set to estimate the poses of shoulder, elbow, and wrist joints. Fig. 12 shows that the proposed method can estimate correctly the locations of the joints even when the body parts of a person are occluded by other people. Fig. 13 (left row) shows failure cases on FLIC dataset of some body poses.

C. *Influence of the Correction Network (CNet)*

From Table III, we can see how the CNet model influences on the performance of pose estimation. The table shows the pose estimation accuracy of each hourglass of the refinement

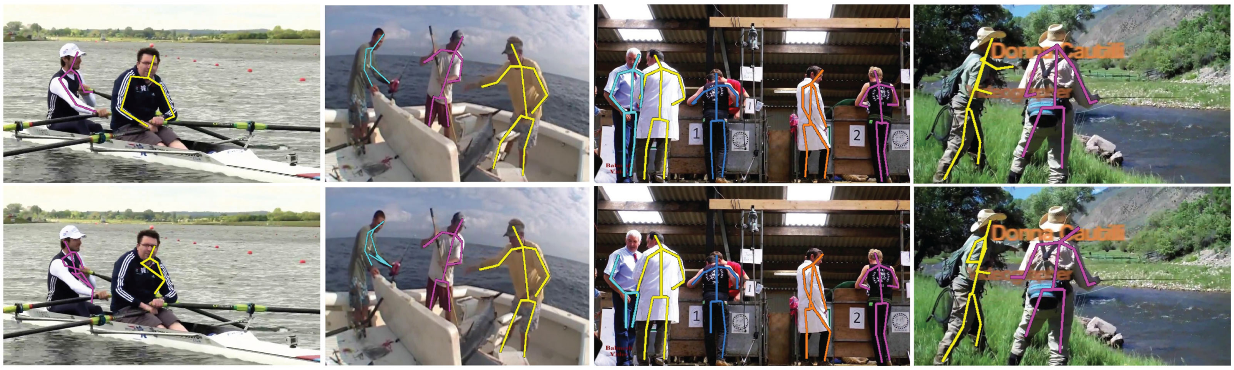


Fig. 10. Visual multi-person pose estimation comparison on images from MPII test set. Down: Results generated by [18]. Top: Results generated by our proposed Hybrid Refinement-Correction method.

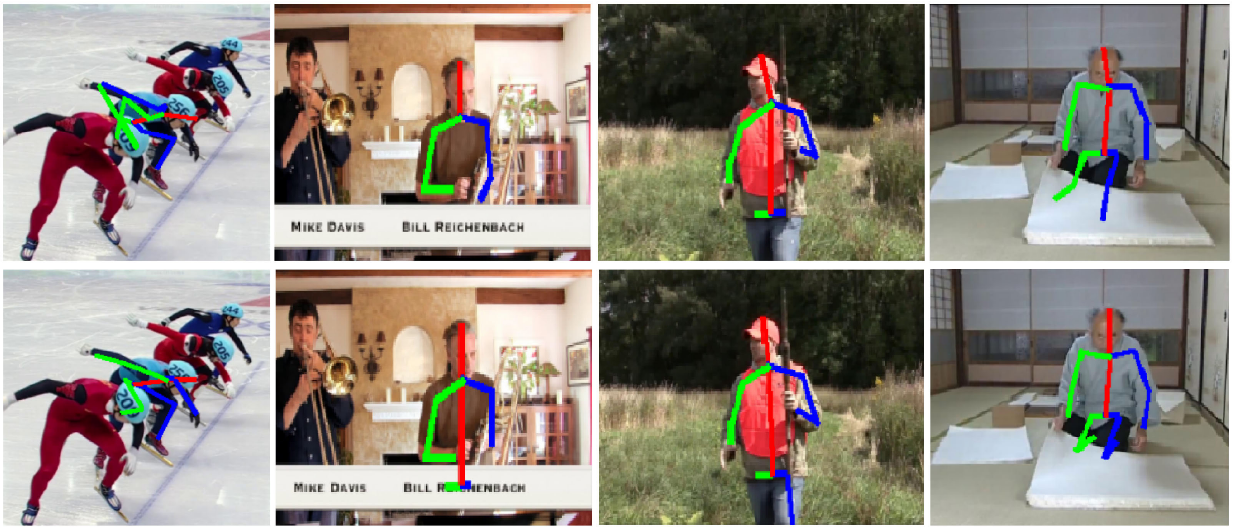


Fig. 11. Visual single-person pose estimation comparison on images from MPII test set. Left: Results generated by [18]. Right: Results generated by our proposed Hybrid Refinement-Correction method.

TABLE II
SINGLE-PERSON POSE ESTIMATION COMPARISON RESULTS ON FLIC TEST SET (PCKH@0.2) USING OUR TRAINED MODEL ON MPII

Method	Elbow	Wrist
Sapp and Taskar [33]	76.5	59.1
Toshev and Szegedy [26]	92.3	82.0
Tompson et al. [25]	93.1	89.0
Chen and Yuille [22]	95.3	92.4
Newell et al. [18]	99.0	97.0
Our proposed (Hybrid-Pose)	**96.5**	**93.2**

model RNet before and after using the hourglasses of the correction model CNet. The results in the second column are obtained by performing the fusion operation between each hourglass *R-HG*$_i$ of the refinement network with all the hourglasses of the correction network (*C-HG*$_1$,...,*C-HG*$_4$), except the last row where the operation is performed between all the hourglasses of the refinement network and all the hourglasses of the correction network. The third column shows the results without involving the correction network. We notice that the pose estimation accuracy after using the correction model is better in all cases. Furthermore, the performance of involving the correction network with all the hourglasses of the refinement network is better than using it with only one hourglass at the time, which reflects the positive influence of the correction network despite the small percentage improvement since pose estimation is a challenging problem and even the improvement by state-of-the-art methods sometimes is less than 1%. The results also give us a sign that hourglasses of various depths can improve the accuracy because of the variation in multi-scale structures, and examining various more depths can improve the results further.

D. Processing Time

The computation complexity is subject to the hourglass depth, the number of the residual modules used in the hourglass, and the total number of hourglasses. A model with more number

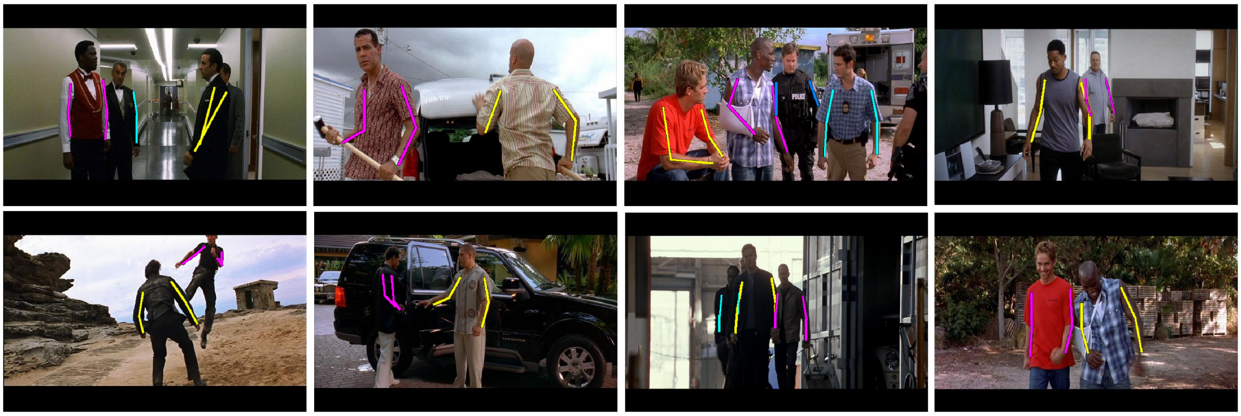


Fig. 12. Upper body pose estimation (shoulder, elbow, and wrist) results on images from FLIC validation set.

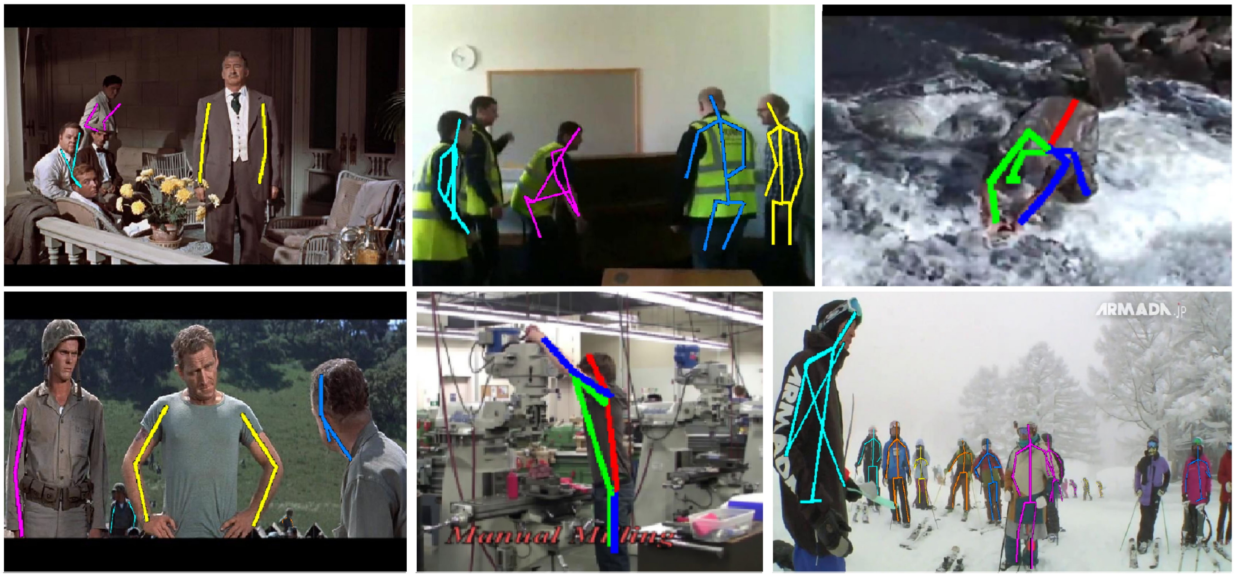


Fig. 13. Pose estimation failure cases on MPII and FLIC images.

TABLE III
COMPARISON BETWEEN THE PREDICTIONS IN CASES OF THE PRESENCE AND THE ABSENCE OF THE CNET MODEL. THE RESULTS ARE OBTAINED ON MPII VALIDATION SET (PCKH@0.5)

C-Net / R-Net	C-HG1 , . . . , C-HG4	Without CNet
R-HG1	**88.8**	85.0
R-HG2	**89.7**	88.1
R-HG3	**90.1**	89.1
R-HG4	**90.2**	89.4
R-HG1 , . . . , R-HG4	**90.4**	89.2

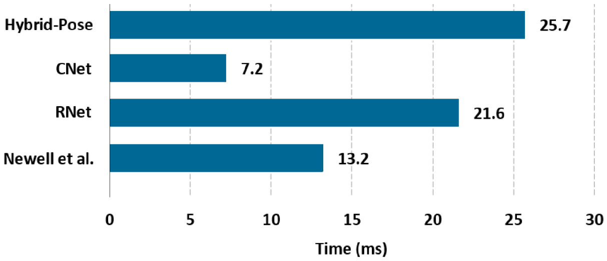


Fig. 14. Processing time required for single-person pose estimation (One forward pass of the model) in milliseconds.

of deep hourglasses performs better but requires more computation resources. In Fig. 14, we present the processing time of a forward pass of the same image using the refinement model RNet, the correction model CNet and the hybrid fusion operation $Fus(RNet, CNet)$ compared with the processing time of the pre-trained model [18]. The comparison shows that the processing time of our hybrid-pose almost equals the sum of the processing time of the RNet model and the CNet model together. Although the heatmaps fusion of the two models generates better results than using a single model, our approach is efficient in term of processing time when using only the CNet model when we look for fast processing with relatively less accuracy (Fig. 7 (C-HG4)).

V. Conclusion

This paper presented a method for single-person pose estimation based on two ConvNet models. One model is used for pose refinement, and the other model is used for pose correction. The two models are designed with different architectures from each other. The correction network learns features that cannot be learned by the refinement network due to its diverse types of hourglasses. The results showed that the fusion of the features generated from the two models improves the accuracy even in difficult body poses. The performance of the proposed method indicates that involving different ConvNet architectures can improve accuracy because of the feature diversity and the wide range of possibilities for joints locations. Future work can focus on improving accuracy by investigating different ConvNet architectures to provide more diverse features for pose correction.

References

[1] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *Int. J. Comput. Vision*, vol. 61, no. 1, pp. 55–79, 2005.

[2] M. A. Fischler and R. A. Elschlager, "The representation and matching of pictorial structures," *IEEE Trans. Comput.*, vol. C-22, no. 1, pp. 67–92, Jan. 1973.

[3] M. Sun and S. Savarese, "Articulated part-based model for joint object detection and pose estimation," in *Proc. Int. Conf. Comput. Vision*, 2011, pp. 723–730.

[4] Y. Tian, C. L. Zitnick, and S. G. Narasimhan, "Exploring the spatial hierarchy of mixture models for human pose estimation," in *Proc. Eur. Conf. Comput. Vision*, 2012, pp. 256–269.

[5] M. Dantone, J. Gall, C. Leistner, and L. V. Gool, "Human pose estimation using body parts dependent joint regressors," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2013, pp. 3041–3048.

[6] X. Lan and D. P. Huttenlocher, "Beyond trees: Common-factor models for 2D human pose recovery," in *Proc. IEEE Int. Conf. Comput. Vision*, 2005, vol. 1, pp. 470–477.

[7] Y. Wang and G. Mori, "Multiple tree models for occlusion and spatial constraints in human pose estimation," in *Proc. Eur. Conf. Comput. Vision*, 2008, pp. 710–724.

[8] V. Ramakrishna, D. Munoz, M. Hebert, J. Andrew Bagnell, and Y. Sheikh, "Pose machines: Articulated pose estimation via inference machines," in *Proc. Eur. Conf. Comput. Vision*, 2014, pp. 33–47.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Advances Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.

[11] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 2818–2826.

[12] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, 2017, pp. 4278–4284.

[13] W. Chu, H. Xue, C. Yao, and D. Cai, "Sparse coding guided spatiotemporal feature learning for abnormal event detection in large videos," *IEEE Trans. Multimedia*, vol. 21, no. 1, pp. 246–255, Jan. 2019.

[14] D. Guo, W. Li, and X. Fang, "Fully convolutional network for multiscale temporal action proposals," *IEEE Trans. Multimedia*, vol. 20, no. 12, pp. 3428–3438, Dec. 2018.

[15] A. Tejero-de-Pablos, Y. Nakashima, T. Sato, N. Yokoya, M. Linna, and E. Rahtu, "Summarization of user-generated sports video by using deep action recognition features," *IEEE Trans. Multimedia*, vol. 20, no. 8, pp. 2000–2011, Aug. 2018.

[16] S. Chen, C. Zhang, and M. Dong, "Deep age estimation: From classification to ranking," *IEEE Trans. Multimedia*, vol. 20, no. 8, pp. 2209–2222, Aug. 2018.

[17] Y. Li, "A deep spatiotemporal perspective for understanding crowd behavior," *IEEE Trans. Multimedia*, vol. 20, no. 12, pp. 3289–3297, Dec. 2018.

[18] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 483–499.

[19] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 4724–4732.

[20] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vision*, 2016, pp. 21–37.

[21] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, "Human pose estimation with iterative error feedback," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 4733–4742.

[22] X. Chen and A. L. Yuille, "Articulated pose estimation by a graphical model with image dependent pairwise relations," in *Proc. Advances Neural Inf. Process. Syst.*, 2014, pp. 1736–1744.

[23] W. Ouyang, X. Chu, and X. Wang, "Multi-source deep learning for human pose estimation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 2337–2344.

[24] T. Pfister, J. Charles, and A. Zisserman, "Flowing ConvNets for human pose estimation in videos," in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 1913–1921.

[25] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 648–656.

[26] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 1653–1660.

[27] L. Dong, X. Chen, R. Wang, Q. Zhang, and E. Izquierdo, "ADORE: An adaptive holons representation framework for human pose estimation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2803–2813, Oct. 2018.

[28] L. Ke, M.-C. Chang, H. Qi, and S. Lyu, "Multi-scale structure-aware network for human pose estimation," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 731–746.

[29] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 1290–1299.

[30] W. Tang, P. Yu, and Y. Wu, "Deeply learned compositional models for human pose estimation," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 197–214.

[31] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 472–487.

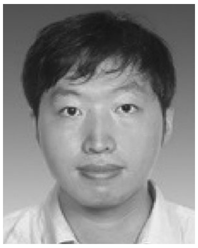
[32] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 5686–5696.

[33] B. Sapp and B. Taskar, "MODEC: Multimodal decomposable models for human pose estimation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2013, pp. 3674–3681.

[34] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 3686–3693.

[35] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2011, pp. 1385–1392.

Aouaidjia Kamel received the M.Eng. degree in computer science from the Abbès Laghrour University of Khenchela, Algeria, in 2009, and the Ph.D. degree in computer science from Shanghai Jiao Tong University, Shanghai, China, in 2019. He is currently with the Visual Media and Data Management Laboratory, Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. He is also a researcher with the Space Techniques Center, Algerian Space Agency, Arzew, Oran, Algeria. His research interests include machine learning, understanding human behaviour, and remote sensing.



Bin Sheng received the B.A. degree in english and the B.Eng. degree in computer science from the Huazhong University of Science and Technology, Wuhan, China, and the M.Sc. degree in software engineering from the University of Macau, Taipa, Macau, and the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, Shatin, Hong Kong. He is currently an Associate Professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. He is an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. His current research interests include image/video processing, virtual reality, machine learning, and computer graphics.

Ping Li (Member, IEEE) received the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, Shatin, Hong Kong. He is currently a Research Assistant Professor with The Hong Kong Polytechnic University, Kowloon, Hong Kong. His current research interests include image/video stylization, artistic rendering and synthesis, and creative media. He has one image/video processing national invention patent, and has excellent research project reported worldwide by *ACM TechNews*.

Jinman Kim received the B.S. (Hons.) and Ph.D. degrees in computer science from The University of Sydney, Sydney, Australia. Since 2006, he has been a Research Associate with the leading teaching hospital, the Royal Prince Alfred. From 2008 to 2012, he was an ARC Post-Doctoral Research Fellow, one year leave from 2009 to 2010 to join the MIRALab Research Group, Geneva, Switzerland, as a Marie Curie Senior Research Fellow. Since 2013, he has been with the School of Information Technologies, The University of Sydney, where he was a Senior Lecturer, and became an Associate Professor in 2016. His current research interests include medical image analysis and visualization, computer aided diagnosis, and telehealth technologies.

David Dagan Feng (Fellow, IEEE) received the M.Eng. degree in electrical engineering and computer science (EECS) from Shanghai Jiao Tong University, Shanghai, China, in 1982, the M.Sc. degree in biocybernetics, and the Ph.D. degree in computer science from the University of California, Los Angeles (UCLA), Los Angeles, CA, in 1985 and 1988, respectively, where he received the Crump Prize for Excellence in Medical Engineering. He is currently the head in the School of Information Technologies, the director in the Biomedical & Multimedia Information Technology Research Group, and the research director in the Institute of Biomedical Engineering and Technology at the University of Sydney, Sydney, Australia. He has published over 700 scholarly research papers, pioneered several new research directions, and made a number of landmark contributions in his field. More importantly, however, is that many of his research results have been translated into solutions to real-life problems and have made tremendous improvements to the quality of life for those concerned. He has served as the chair in the International Federation of Automatic Control (IFAC) Technical Committee on Biological and Medical Systems, has organized/chaired over 100 major international conferences/symposia/workshops, and has been invited to give over 100 keynote presentations in 23 countries and regions. He is a fellow of the IEEE and Australian Academy of Technological Sciences and Engineering.