IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 26, 2020, accepted July 15, 2020, date of publication July 20, 2020, date of current version July 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3010257

Multi-Person Pose Estimation Under Complex Environment Based on Progressive Rotation Correction and Multi-Scale Feature Fusion

GUOHENG HUANG[1], (Member, IEEE), XIAOPING CHEN[1], JUNAN CHEN[1], WEIDA LIN[1], WING-KUEN LING[2], (Senior Member, IEEE), CHI-MAN PUN[3], (Senior Member, IEEE), LIANGLUN CHENG[1], (Senior Member, IEEE), AND ZHUOWEI WANG[1]

[1]School of Computers, Guangdong University of Technology, Guangzhou 510006, China
[2]School of Information Engineering, Guangdong University of Technology, Guangzhou 510006, China
[3]Department of Computer and Information Science, University of Macau, Macau 999078, China

Corresponding authors: Wing-Kuen Ling (yongquanling@gdut.edu.cn), Chi-Man Pun (cmpun@umac.mo), Lianglun Cheng (llcheng@gdut.edu.cn), and Zhuowei Wang (wangzhuowei0710@163.com)

This work was supported in part by the National Nature Science Foundation of China under Grant 61702111, in part by the National Key Research and Development Program of China under Grant 2017YFB1201203, in part by the Guangdong Provincial Key Laboratory of Cyber-Physical System under Grant 2016B030301008, in part by the National Natural Science Foundation of Guangdong Joint Fund under Grant U1801263, in part by the National Nature Science Foundation of China-Guangdong Joint Fund under Grant 83-Y40G33-9001-18/20, in part by the National Natural Science Foundation of Guangdong Joint Fund under Grant U1701262, in part by the Key-Area Research and Development Program of Guangdong Province under Grant 2019B010153002, Grant 2018B010109007, and Grant 2019B010109001, and in part by the "Blue Fire Plan" (Huizhou) Industry-University-Research Joint Innovation Fund 2017 Project of the Ministry of Education under Grant CXZJHZ201730.

ABSTRACT The research of multi-person pose estimation has been largely improved recently. However, multi-person pose estimation in complex environments is still challenging. For example, the following two situations cannot be handled well by existing pose estimation methods: first, there are pedestrians that are not upright or even inverted in the image, and pedestrians of different scales appear in the same image. To solve these problems, the Progressive rotation correction module (PRCM) and Scale-invariance module (SIM) based on multi-scale feature fusion are proposed. First of all, the PRCM was proposed to address the situation where pedestrians appear rotated or even inverted in the image. This module is divided into three stages, with the aim of gradually correcting the inverted human to an upright one. Besides, SIM is designed to handle multi-scale problems. In this module, dilated convolutions with different receptive field are used to extract multi-scale information. Then, the extracted multi-scale features (different semantic information in different feature maps) will be fused to solve the multi-scale problem. The experimental results show that our algorithm can reach an AP value of 72.0% when tested on the COCO2017 dataset. Demonstrates that the proposed method is superior to state-of-the-art methods.

INDEX TERMS Pose estimation, rotation invariance, multi-scale feature fusion, dilated convolution.

I. INTRODUCTION

Human pose estimation has always been a challenging research area in computer vision. The purpose of human pose estimation is to automatically capture the position of the key points of human body. It is the foundation of human-computer interaction and action recognition. Especially in the past ten years, research on human pose estimation has become more active. In the first place, Graphic structure or graphic model technology is generally used in traditional human pose estimation methods [1], [2]. More specifically, the key point estimation problem of single person is expressed as a tree structure or a graphical model problem, and the key point positions are predicted based on the manual features.

Recent works mostly rely on the development of convolutional neural network, which largely improve the performance of pose estimation [3]–[10]. Pose estimation is divided into single-person pose estimation and multi-person pose estimation, where single-person pose estimation is based on a single human body predicting key points, and multi-person

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

pose estimation needs to further identify the key points of everyone within an image. At present, most of single-person pose estimation methods are based on deep learning frameworks. Among them, Tompson *et al.* first predicted the coordinates of each key point through a convolutional neural network, and used the structural relationship between key points of the human body to combine Markov random field to optimize the prediction result [3]. Tomas *et al.* predicted the heatmap of each key point by convolutional neural network to return to the key point of the human body, which greatly improved the single coordinate point compared with the regression. This network is completed based on successive steps of aggregation and upsampling to produce the final prediction result [4]. Newell *et al.* proposed the structure of Hourglass in 2016, which was first designed for single-person pose estimation (now it has been generalized to multi-person pose estimation) [5]. By repeating bottom-up and top-down and jointly supervising the intermediate results, they can make good use of the Spatial relationships between different parts of the body. In addition, Wei *et al.* designed a Convolutional Pose Machine (CPM) which is a multi-stage architecture, first output a rough pose estimation result, and then continue to refine this result in following stages [6].

Although single-person pose estimation has achieved good results, scenarios with multi-person are more common in practical applications. Therefore, multi-person pose estimation is gradually being more widely studied. The approach of multi-person pose estimation is mainly divided into two categories: bottom-up approaches and top-down approaches. The bottom-up approach first detects all the key points of the human body and then combines them into a complete skeleton. Cao *et al.* proposed Partial Affinity Fields (PAFs) to learn to associate body parts with individuals in images. This architecture aims to learn part locations and their associations together through two branches of the same sequential prediction process [7]. In addition, Newell *et al.* proposed a network that could simultaneously output detection results and group allocation information, which could determine which joint belongs to which person [8]. The top-down approach interprets the process of detecting key points as a two-stage pipeline, that is, first locating and cropping all persons from the image, and then solving the single-person pose estimation problem. GlobalNet uses the Cascade Pyramid Network (CPN) to detect points that are easier to detect, while RefineNet is used to detect points that are more difficult to detect [9]. He *et al.* proposed a flexible network structure Mask R-CNN, which added branches to the mask that carried the key points of the human body [10]. In the top-down multi-person pose estimation method, the first step is to detect person from the image and then perform single-person pose estimation [9], [11], [12]. Where, person detection is mainly through the R-CNN family [13]–[15] and YOLO [16], [17]. Our work is top-down multi-person pose estimation, using YOLOv3 to detect the human body. Besides, as mentioned, the stacked Hourglass Network was also applicable to multi-person pose estimation [5].

Although pose estimation has made great progress, there are still some issues that remain unresolved (as shown in FIGURE 1). First, the existing method cannot handle the case where the person in the image is rotated or inverted. In addition, the distance between each person in the photo and the camera may vary widely. However, the existing methods cannot effectively solve the problem of large differences in scales among individuals in the same image.

(a) (b)

FIGURE 1. Example of multi-person pose estimation in complex situations: (a) shows three persons in a rotated state and (b) shows two persons with different scales in the same image.

Aiming at the difficult issues mentioned above, we propose a novel pose estimation scheme with rotation-invariance and scale-invariance in this paper. In summary, the contributions are as follows:

1 In order to solve the problem of rotated or inverted of the human body in the image, we propose a novel Progressive rotation correction module (PRCM). This module contains three stages that progressively correct a rotated or inverted image to an upright position.
2 In order to solve the multi-scale problem in pose estimation, we propose a Scale-invariant module (SIM) based on multi-scale feature fusion. This module includes two parts: multi-scale learning by different receptive field and multi-scale feature fusion. Feature extraction is performed by adding different dilated convolutions to different feature maps, and then fuses feature maps of different scales.
3 Our network has two modules and can be trained jointly. The loss of PRCM is the Euclidean distance between the predicted value of the regression box and the ground-truth. The loss of SIM is the Euclidean distance between the coordinates of the predicted key points and the ground-truth. The cumulative sum of the loss of these two modules is taken as the loss of the entire network.
4 Based on the proposed algorithm, we achieve better results on the COCO key points benchmark, with average precision at 72.0% on the COCO2017 dataset, which is a 0.6% relative improvement compared with 71.4%.

In the remainder of this paper, we first discuss related works about rotation-invariant and multi-scale feature extraction in section II. We then introduce the details of proposed method section III. In section IV, extensive experiments are performed to evaluate the performance of proposed method.

Finally, conclusions and future works are summarized in section V.

II. RELATED WORK

A. STRATEGIES OF ROTATION-INVARIANCE IN DEEP NEURAL NETWORKS

In multi-person pose estimation, there are still some complex scenes that are difficult to handle, such as pedestrians tilting or even upside down in an image. Although CNN is used to process distorted and rotated images, it cannot handle well of an image with person that rotates at a large angle. Once upon a time, data augmentation is used in most methods to deal with images of rotated or slanted person, but this method could only solve small-angle rotations [9]. Nowadays, more and more rotation-invariance methods are being considered in deep learning networks. For example, Spatial Transformation Network (STN) allows spatial transformation of data in the network explicitly [18]. This differentiable module can be inserted into the existing convolutional architecture, enabling the neural network to actively transform feature maps in space to achieve the effect of processing rotation. Deformable convolution is the improved version of STN, which enhances the spatial information modelling capabilities of the current CNN network through a variable convolution structure, thereby achieving the effect of processing rotation [19]. In order to extract the rotation-invariant feature, Cheng *et al.* added a rotation-invariance fully connected layer to the network, and optimized a new objective function by adding a regular constraint term to ensure that training samples share similar features before and after rotation [20]. Daniel *et al.* applied a rich, parameter-efficient and fixed computational complexity representation, showing that deep feature maps within the network encode complicated rotational invariants [21]. Welling *et al.* introduced Group Equivariant Convolutional Neural Networks (G-CNNs) [22]. Among them, a network with isomorphism under a specific transformation (rotation, translation, etc., which can also be expressed as a special group) is proposed. They conducted experiments on MNIST and CIFAR data with rotation transformation, and confirmed that the rotation group CNN can resist rotation better. In addition, Shi *et al.* proposed a Progressive Calibration Network (PCN) to perform rotation-invariant face detection in a coarse-to-fine manner [23].

Deformable convolution can handle with rotating objects by a well-designed deformable convolution structure. This structure deals with small-angle rotations well, but it does not handle large-angle rotations, such as the inverted case [19]. Shi *et al.* proposed a progressive calibration network (PCN) to complete the face detection process from coarse to fine [23]. In the early stage of the algorithm, a rough direction estimation was performed, and then an accurate direction adjustment was performed. In this way, the network can complete end-to-end multi-directional face detection. PCN consists of three stages, the first stage is to select a portion of the candidate faces from the input image, remove the non-face candidate box and place the Rough rotation of the candidate box. The second stage further distinguishes more accurately between faces and non-faces, regresses the bounding box, and calibrates the face candidates. The third stage makes the final decision easily, accurately determines whether it is a face or not, and regresses the bounding box. Inspired by PCN, we design a Progressive rotation correction module to progressively (with three stages) correct a human body in a rotated or inverted situation to an upright state in pose estimation. The three stages of this module can gradually correct an inverted person object to an upright state.

B. ACCURATE MULTI-SCALE PEDESTRIAN DETECTION

In the same image, the distance between each person being photographed and the lens may be very different, which will cause a large deviation in the scale of each person in the captured image. In order to solve the problem caused by multi-scale, most methods apply multi-scale feature fusion strategies. Especially, earlier multi-scale fusion methods usually appeared in object detection tasks, and convolution kernels of different sizes were designed to fit different sized objects in the image structure. These methods are commonly used to detect objects at different scales [24]. Redmon *et al.* used upsampling and fusion methods to fuse information at 3 scales and independently perform detection on the fusion feature maps at multiple scales [25]. Because the lower scale has less feature semantic information, but accurate target location information; the higher scale has richer feature semantic information, but the target location information is coarse. Therefore, this multiscale feature fusion approach is more effective in detecting small targets. In addition, Fisher *et al.* proposed a Feature Pyramid Network (FPN) to deal with the scale change of body parts [26]. Besides, Newell's Stacked Hourglass Networks can learn the local features of key points through a multi-scale receptive field mechanism [5]. This Hourglass module is designed to capture the local information contained in images at different scales, while the final pose estimation requires a consistent understanding of the whole body. The GlobalNet module were designed by Chen *et al* [9]. This module fuses feature maps with different scales of different receptive fields at different layers to handle multi-scale problems. Besides, Sun *et al.* have proposed a High-Resolution Network (HRNet) [27], which consists of parallel high- and low-resolution subnetworks, and this networks exchange information repeatedly between multi-resolution subnets (multi-scale fusion). They repeatedly perform multi-scale fusion so that each high-and low-resolution representation can repeatedly receive feature map information from other scales. Ronneberger *et al.* proposed a method consisted of a contracting path to capture the context, and a symmetric expanding path that enabled precise localization [28]. This structure is a combination of multi-channel convolution and FPN structure, and the output of the feature extraction part will be fused in each subsequent layer. Aiming at the problem that it is difficult to identify smaller objects and larger objects in an image,

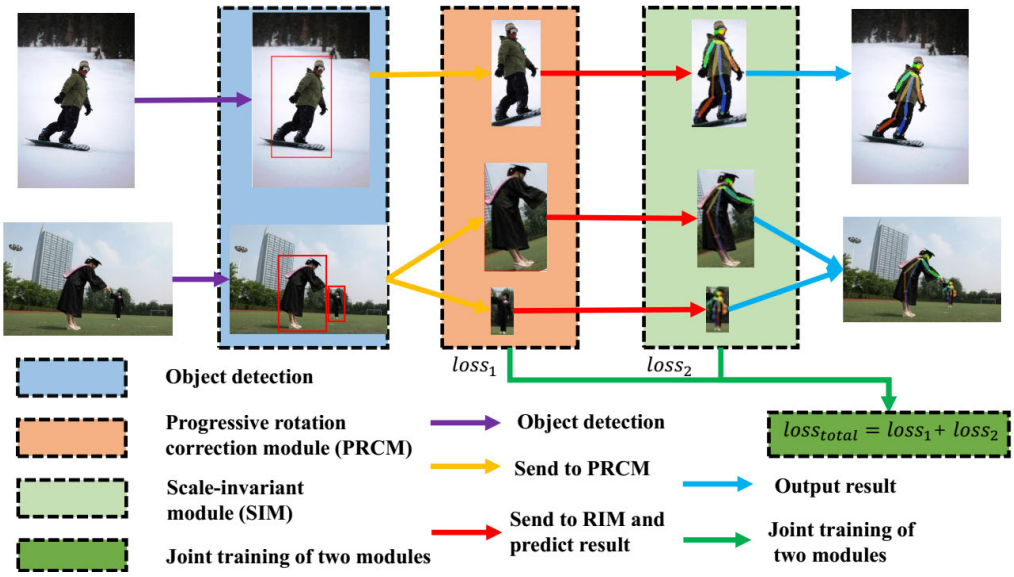


FIGURE 2. The whole architecture of our network: First, a single person is detected by the object detection network, and then the single person image is sent to the Progressive rotation correction module to correct the position. The corrected image is sent to the Scale-invariant module for feature extraction. At the same time, two modules can be trained jointly. Finally, the final result is output.

Singh *et al.* proposed a novel training scheme called Scale Normalization for Image Pyramids (SNIP) which selectively back-propagates the gradients of object instances of different sizes as a function of the image scale [29].

As mentioned, many methods design different receptive fields for different feature maps to learn multi-scale features in multi-scale problems. Dilated convolution can enlarge the convolutional kernel with original weights by performing convolution at sparsely sampled locations, thus increasing the receptive field size without additional parameter cost [26]. Dilated convolution is widely used for semantic segmentation to extract context information by expanding the receptive field [30], [31]. The dilated convolution can expand the receptive field, without introducing additional parameters, and can capture multi-scale contextual information. A large convolution kernel is used in Convolutional Pose Machines (CPM) networks to extend the receptive field to obtain context information [32]. Furthermore, Deformable Convolution is a more general convolution operator by adaptively learning 2D offsets [19]. Li *et al.* found that different receptive fields have different effects on different scales [33]. They designed the Trident Networks that use different dilated convolutions to adapt to objects of different scales.

Inspired by these previous studies, for multi-scale problems in images, we can expand the receptive field method in large-scale images to obtain key points information, and similarly reduce the receptive field appropriately on small scales. Then the feature maps of different scales are fused to solve the multi-scale problem.

III. OUR APPROACH

In this section, we introduce the novel algorithm for multi-person pose estimation in complex environments in details. Like most networks, our method is a top-down approach. First of all, each person is detected by YOLOv3 from each input image. In the second step, the detected single person image is sent to a Progressive rotation correction module (PRCM), and the rotated human body is gradually corrected to an upright human body. In the third step, the corrected human image is sent to the Scale-invariant module (SIM) for further feature extraction. In this module, multi-scale information is learned through convolution kernels of different receptive fields, and multi-scale information is fused to deal with multi-scale problems. In addition, our two modules can be jointly trained, and the parameters of the network are optimized by accumulating the losses of these two modules and backpropagating. Finally, the key points of the human body are predicted through the trained network. The overall flow of the whole algorithm is shown in FIGURE 2.

A. PROGRESSIVE ROTATION CORRECTION MODULE

At the beginning of the network, each person is detected from an image with by using an object detection algorithm (e.g. YOLOv3). The detected bounding box of each single person is resized to 384×288, and then sent to the Progressive rotation correction module (PRCM) to correct the rotating human body. Especially, the proposed PRCM consists of three stage that gradually calibrate the Rotation-Of-Angle (ROA) of candidates to upright. Here, we select the centre point of the candidate box. Then make a straight line perpendicular to the image through the centre point as the central axis of rotation. In the following, ROA is based on the central axis as a rotation.

After an image is input, all candidates are generated according to the sliding window principle. Candidates are continuously trained by PRCM to regress the person

bounding box and ROA. Actually, calibration means rotating the body at the ROA angle to get an upright body. The first stage, in PRCM-1, candidates are identified from images and candidates are calibrated from bottom to top by correcting the ROA angle range from [−180°, 180°] to [−90°, 90°]. In the second stage, in PRCM-2, the ROAs of candidates are further distinguished and calibrated to the vertical range of [−45°, 45°], thereby reducing the ROA range by half again. In the third stage, PRCM-3 can accurately and quickly identify the candidate and correct the candidate to an upright state. The proposed PRCM is shown in FIGURE 3.

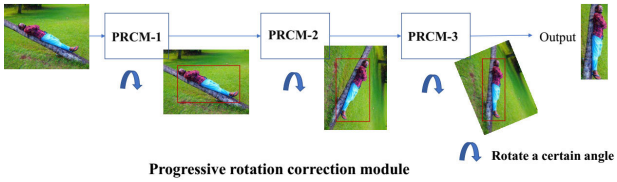


FIGURE 3. An overview of the Progressive rotation correction module: Our PRCM gradually calibrates the ROA direction of each candidate to be upright to better regress to the bounding box. The person in the image is rotated 60° to the left. In the first stage, no rotation is needed because the candidate is already an angle in the range [−90°, 90°]. In the second stage, the target candidate needs to rotate 90° to the right to correct it to an angle in the range [−45°, 45°]. In the third stage, the candidate is rotated to the left by a small angle to correct it to an upright state.

1) PRCM-1 IN FIRST STAGE

For each input single person image, it can be expressed as x. PRCM-1 has two goals: regression bounding box and rough correction, as shown in Equation 1:

$$[t, g] = F_1(x) \tag{1}$$

where F_1 is a detector with a small CNN in the first stage. Variable t is the score representing the bounding box regression prediction, and g is the ROA rotation angle score.

The second objective attempts to make the regression a fine bounding box, as shown in Equation 2:

$$L_{reg}(t, t^*) = S(t - t^*) \tag{2}$$

where S represents loss, t and t^* represent the ground-truth regression results and the predicted results. The regression of the bounding box can be expressed as Equation 3:

$$\begin{aligned} t_w &= w^*/w, \\ t_h &= h^*/h, \\ t_a &= (a^* + 0.5w^* - a - 0.5w)/w^*, \\ t_b &= (b^* + 0.5h^* - b - 0.5h)/h^*, \end{aligned} \tag{3}$$

where a and b denote the top-left coordinates of the bounding box, w and h denote its width and height. Here, a^* and a are for the ground-truth box and predicted box respectively.

The second goal is to roughly predict the candidate's orientation in a binary classification as follows:

$$L_{cal} = y \log g + (1 - y) \log(1 - g) \tag{4}$$

where y is equal to 1, it means that the input image x is upright, and when y is equal to 0, it means that the person body is upside down.

Overall, the objectives of the first stage of PRCM-1 are defined as:

$$loss_1 = \min_{F_1} L = \lambda_{reg} \cdot L_{reg} + \lambda_{cal} \cdot L_{cal} \tag{5}$$

where λ_{reg}, λ_{cal} are parameters to balance the loss of first objective and second objective.

After optimizing Equation 5, PRCM-1 can be used to filter some candidates. For the remaining candidate boxes, the new bounding box of the regression is updated according to PRCM-1. Then, the updated candidate frame is rotated based on the predicted rough ROA angle. The ROA angle predicted in the first stage is expressed as θ_1, which is calculated by the following formula:

$$\theta_1 = \begin{cases} 0^\circ, & g > 0.5 \\ 180^\circ, & g \leq 0.5 \end{cases} \tag{6}$$

Specifically, $\theta_1 = 0^\circ$ means that the candidate is facing up, so there is no need to rotate, otherwise $\theta_1 = 180^\circ$ means that the candidate is facing down, and it needs to be rotated 180°. In this way, the range of the ROA angle is reduced from [−180°, 180°] to [−90°, 90°].

2) PRCM-2 IN SECOND STAGE

The PRCM-2 in the second stage is similar to the PRCM-1 in the first stage, and further returns to the bounding box and calibrates the candidates. Different from the first stage, the rough prediction of the ROA angle range at this stage is ternary classifications, that is, [−90°, 45°], [−45°, 45°] and [−45°, 90°]. Predict the ROA angle and perform rotation calibration in the second stage, as shown in Equation 7:

$$\begin{aligned} id &= \arg\max g_i, \\ \theta_1 &= \begin{cases} -90^\circ, & id = 0 \\ 0^\circ, & id = 1 \\ 90^\circ, & id = 2 \end{cases} \end{aligned} \tag{7}$$

where g_0, g_1 and g_2 are the predicted ternary classification scores. The candidates are rotated -90°, 0° or 90° respectively. After the second stage, the range of the ROA angle is reduced from [−90°, 90°] to [−45°, 45°].

3) PRCM-3 IN THIRD STAGE

After the second stage, all candidates are calibrated to a ROA angle range of [−45°, 45°]. Therefore, the third stage of PRCM-3 can easily correct the candidate to an upright state because the angle of the offset is relatively small, and the regression bounding box can be accurately determined. Because the ROA angle has been reduced to a smaller range in the previous stage, PRCM-3 attempts to directly return the candidate's exact ROA angle, rather than a rough orientation.

Finally, the candidate's ROA angle can be obtained by accumulating the predictions of the three stages, expressed

as θ, as follows:

$$\theta_{ROA} = \theta_1 + \theta_2 + \theta_3 \tag{8}$$

The rotation angle of the candidate is obtained by the sum of the ROA angles predicted in the three stages, that is, $\theta_{ROA} = \theta_1 + \theta_2 + \theta_3$. In particular, θ_1 has only two values 0° or 180°, θ_2 has only three values 0°, 90°, or −90°, and θ_3 is a value in the range [−45°, 45°]. We also provide some examples for calculating the ROA angle, as shown in FIGURE 4.

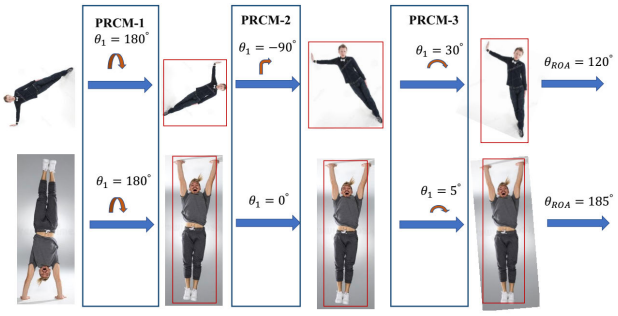


FIGURE 4. Three stages of calculating ROA: Two rotated and inverted images, after three stages of PRCM, are corrected to an upright state.

4) ARCHITECURE OF PRCM

As mentioned, the proposed PRCM consists of three stages. The input image is 80 × 60 pixels. The output of each stage is the ROA angle and the coordinates of the regression box. In PRCM-1, a 3 × 3 convolution kernel is used for convolution, and the stride is 2, each time the convolution is performed, the size of the feature map is reduced by half, and the dimension is doubled. At the end of the network, a fully connected operation is used to output the ROA and bounding box regression. Similarly, in PRCM-2 and PRCM-3, a convolution structure similar to PRCM-1 is used, except that the output of the fully connected layer is different. The PRCM network structure is shown in FIGURE 5.

In order to better train the PRCM, we added two attributes to the COCO dataset label, one is the ROA angle θ, and the other is the coordinates of the human body's bounding box after corrected to upright. When creating the dataset, we first tested the training images with baseline to find out the larger rotated and inverted bounding boxes. At the same time, mark the ROA of images with small rotation angles as 0. In the training process, we only need to return the coordinates of the human body's bounding box and the ROA. By obtaining these two attributes through the trained model, the inverted human body in the image can be corrected to an upright human body.

B. SCALE-INVARIANT MODULE BASED ON MULTI-SCALE FEATURE FUSION

In this section, we will introduce the Scale-invariant module (SIM) in details to solve the problem of large differences in object scale in the image. After the original image is sent to the Progressive rotation correction module (PRCM),

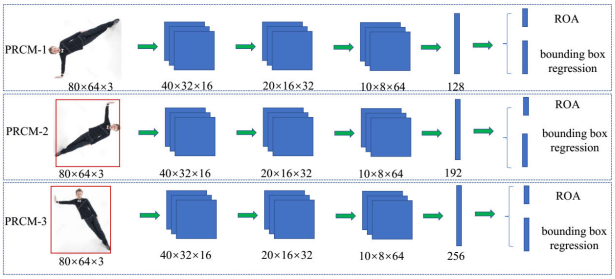


FIGURE 5. Structure of PRCM: The detailed CNN structure of the three stages in the proposed PRCM method. The input is the original image. In order to show the rotation process more intuitively, we rotated the input image by ROA. The output of the network is the of the ROA and the bounding box regression.

the corrected upright image will be output. Then, the corrected image is sent to the SIM for further feature extraction to deal with multi-scale problems.

Our SIM contains two parts, one is to learn multi-scale information using novel dilated convolution, and the other one is multi-scale feature fusion. On one hand, the information of feature maps of different scales is different. In large scale feature maps have more detailed information, while in small scale feature maps the information is more abstract. Therefore, we have designed a novel dilated convolution that uses different dilated convolutions for feature maps of different scales. In different feature maps, multi-scale information is learned through the adaptation of different receptive fields. On the other hand, the features of each part of the human body in pose estimation are not concentrated on the feature map of the last layer, and features in different parts may be distributed among feature maps of different scales. Therefore, we need to fuse multi-scale features to improve the detection effect. Regarding the design of the SIM, we use different rate of dilated convolution to obtain different receptive fields for feature map at different scales, as detailed in 1). After learning multi-scale information through different receptive fields, then multi-scale fusion is performed, which is detailed in 2). The specific structure of Scale-invariant module based on multi-scale feature fusion is shown in FIGURE 6.

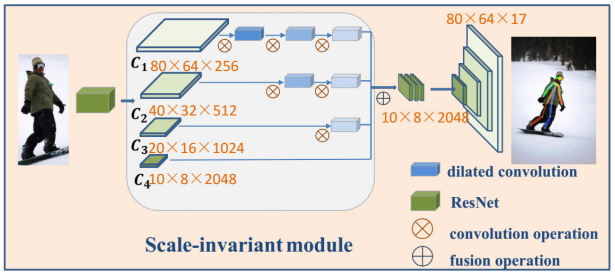


FIGURE 6. Structure of Scale-invariant module (SIM): The corrected single-person image is sent to ResNet for feature extraction. First, multi-scale feature maps are learned by using dilated convolutions of different receptive fields. Then feature maps of different scales are fused. Finally, deconvolution is used to expand the feature map to predict key points.

1) MULTI-SCALE LEARNING BY DIFFERENT RECEPTIVE FIELD

In this process, we use ResNet as the backbone network, and the output of the last residual module is represented as C_1, C_2, C_3 and C_4. In large-scale layers, more effective information is extracted using large receptive fields. Conversely, on a smaller scale, the deeper the network is, the more abstract the extracted semantic information is. If large receptive fields are used in small objects for feature extraction, small objects will be lost. Therefore, we have designed a novel dilated convolution, which can adaptively learn multi-scale information through different receptive fields. The proposed dilated convolution is shown in FIGURE 7.

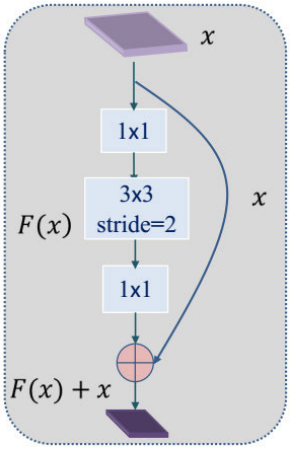


FIGURE 7. Dilated convolution of Scale-invariant module: The input feature map x is convolved with 1× 1, 3 × 3 and 1 × 1 added to the input x. Here, different convolution kernels can be used in dilated convolutions. The output feature map is reduced to half of the original feature.

For feature maps of different scales, we use convolution kernels of different dilated convolution rates to convolve to extract multi-scale information. Multi-scale features can be obtained by using dilated convolution which is a convolution operation on a feature map using an interval convolution kernel. This convolution process can not only expand the receptive field, but also reduce the information loss caused by downsampling.

In this paper, we use the dilated convolution to learn the features of each scale and consider a dilated convolution block defined as:

$$y = F(x) + x \tag{9}$$

where x and y are the input and output vectors of the layers considered. The function $F(x)$ represents the residual mapping to be learned. The operation $F(x) + x$ is performed by a shortcut connection and element-wise addition. In this convolution, first, we use a 1×1 convolution kernel to expand the dimensions, and then use a 3×3 convolution kernel to convolve. Different dilated convolution rates are used in this dilated convolution. The number of input and output channels is the same. The dimensions of the input feature map will be half of the original dimensions.

Our method is to use a convolution kernel of $d = 3$ in the larger feature map C_1, a convolution kernel with $d = 2$ for the medium feature map C_2, and a convolution kernel with $d = 1$ for the small feature map C_3. There, the dimensions of the input and output remain the same, but the feature map shrinks in general, using different dilated convolution rates in the 3×3 convolution. Each time the dilated convolution module is used, the number of channels of the feature map is doubled, but the size of the feature map is reduced to half its original size.

2) MULTI-SCALE FEATURE FUSION

In this section, we introduce the multi-scale feature fusion in detail. Our method is based on the ResNet backbone, and different levels of feature maps are obtained by different downsampling in the four stages of ResNet. The result is the original size, half size, quarter size and one eighth size of the feature map. For the multi-scale problem of people in images, we first feature extracted on different feature maps using convolutional kernels with different dilated convolution rates, and then fused the different scales of feature maps.

As the number of layers of the convolutional neural network deepens, the semantic information is extracted from the bottom to the top layers. For the extraction of key points of the human body, the features extracted at the lower layers of the network are only some contour features. As the network deepens, the extracted features are higher semantic features such as eyes, nose, etc., and finally the network reaches the deepest layer to extract the key points throughout the human body. However, as the network deepens, each layer loses some information, and more information is lost in the last layer. Therefore, we add the feature map of the previous layer to the feature map of that layer, which preserves some information of the previous layer and reduces the loss of that layer Information. For pose estimation, not all features are concentrated on the feature map of the last layer, and different key points may be distributed on feature maps of different scales. If only the feature map of the last layer is used to detect the key points, the detection effect will not be good.

The feature maps at different scales contain different levels of semantic information, and fusing the information from different scale layers can be better adapted to the images at different scales. Using the fused feature map for detection not only improves the accuracy of small object detection, but also does not lose the detection of large object. Therefore, the feature map is fused and then the key points are predicted. The feature maps are defined as C_1, C_2, C_3 and C_4 from large to small. After these feature maps of different sizes are passed through the dilated convolution module, these feature maps are fused. The process of fusion can be expressed as Equation 10.

$$R_{out} = F(C_1 \otimes D_1) \oplus F(C_2 \otimes D_2) \oplus F(C_3 \otimes D_3) \oplus C_4 \tag{10}$$

where R_{out} represents the result after fusion, C represents four different size feature maps, D represents four different convolution kernels. There F represents the convolution result of the current size. The symbol $\otimes$ represents a convolution operation, and $\oplus$ represents a fusion operation.

C. JOINT TRAINING OF TWO MODULES

After the feature extraction of the Scale-invariance module (SIM), the resolution of the feature map at this time is only 10×8. At this time, it is necessary to generate a feature map of 80×64 resolution for key points prediction. We use deconvolution to generate high-resolution feature maps. If we use upsampling directly, some information will be lost, so we use a deconvolution operation, which is equivalent to convolution and upsampling, so that more feature information can be retained.

After using three deconvolution layers, the size of the generated feature map is $80 \times 64 \times 2048$, where 2048 represents the number of channels. Then use a 1×1 convolution to generate k heat maps$\{H_1, H_2, H_3 \ldots H_k\}$, and finally predict the key points in the heat map.

For the calculation Scale-invariant module of loss, we use a Gaussian function to generate a heat map for each key point, and also generate a corresponding heat map for the predicted key points. The loss function is calculated as Equation 11. θ_i represents the position of the real coordinates of the i-th key point, $\hat{\theta}_i$ represents the predicted coordinates. Where, s represents the size of the current image, and k represents k key points in total. Calculate the Euclidean distance from the ground-truth of each key point and the predicted heat map. The loss of SIM can be defined as:

$$loss_2 = \sum_{i=1}^{k} e^{-\frac{\|\hat{\theta}_i - \theta_i\|_2^2}{2s^2k^2}} \tag{11}$$

As mentioned, there are two modules in our network, and the loss of these two modules must be back-propagated. Let's record the loss of the Progressive rotation correction module (PRCM) as $loss_1$ ich already defined in Equation 5, and the loss of the SIM as $loss_2$, then the loss of the entire process is $loss_{total}$, as shown in Equation 12. During the training process, both parts of the loss work, and the parameters are optimized by backpropagation.

$$loss_{total} = loss_1 + loss_2 \tag{12}$$

IV. EXPERIMENT ANALYSIS

In this section, we will discuss more details of our methods according to the experiment results. Our overall pipeline follows the top-down approach for multiple human pose estimation. Firstly, we apply an object detector to generate human proposals. For each proposal, we assume that there is only one main person in the proposed cropped region, and then apply a pose estimation network to generate the final prediction. We first describe the experimental environment and evaluation indicators in section A. In section B, we experimentally verify our proposed Progressive rotation correction module and Scale-invariance module.

A. EXPERIMENTAL ENVIRONMENT AND EVALUATION METRIC

1) DATASET DESCRIPTION

Our network is trained on MS COCO trainval dataset (includes 80000 images and 120000 person instances) and validated on MS COCO minival dataset (includes 500 images). The testing sets includes test-dev set (12000 images). In order to minimize the variance of prediction, we apply a gaussian filter on the predicted heatmaps. Following the same techniques used in [5], we also predict the pose of the corresponding flipped image and average the heatmaps to get the final prediction.

2) EXPERIMENTAL ENVIRONMENT

All proposed models of pose estimation are trained using Adaptive moment estimation (Adam) algorithm with an initial learning rate of 0.0001. Batch normalization is used in our network. Generally, the training of ResNet101 based models takes about 1 day on eight NVIDIA Tesla K80 GPUs. Our models are all initialized with weights of the public-released ImageNet [34] pretrained model.

3) EVALUATION METRIC

Our experiments were evaluated in Object Key Point Similarity (OKS) where OKS define the similarity between different human poses. OKS can be defined as Equation 13.

$$OKS = \frac{\sum_i exp(-d_i^2/2s^2k_i^2)\delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \tag{13}$$

where d_i is the Euclidean distance between the detected key points and the corresponding ground truth, d_i is the v_i is the visibility flag of the ground truth, s is the object scale, and k_i is a per-keypoint constant that controls falloff. We report standard Average Precision (AP) and recall scores: AP^{50} (AP at $OKS = 0.50$), AP^{75}, AP (the mean of AP scores at 10 positions, $OKS = 0.50, 0.55 \ldots 0.90, 0.95$, AP^M for medium objects, AP^L for large objects, and AR are averaged over multiple $OKS = 0.50, 0.55 \ldots 0.90, 0.95$) [27].

B. COMPARISON WITH STATE-OF-THE-ART METHODS

1) COMPARATIVE RESULTS ON PROGRESSIVE ROTATION CORRECTION MODULE

To verify the effectiveness of the Progressive rotation correction module (PRCM), we only add PRCM to the network. Inverted and rotated images are added to the network for training to enhance the ability of PRCM to adapt to complex scenes. There, our method is not data augmentation, but simply correcting the rotated human body in the input image to an upright state, without generating new data. In fact, the proposed algorithm rotates only the bounding box and not the whole image.

Our method is mainly compared with state-of-the-art methods: Mask-RCNN [10], G-RMI [35], RMPE [11] and

MultiPoseNet [36]. At the same time, we specially selected images with a large rotation angle and inverted images for testing. Since the methods such as Mask-RCNN cannot deal with the rotation and inverted situation specially, the detection error occurs in the image of the rotated scene. In contrast, our method can detect complete human key points on rotated and inverted images. Through the PRCM, the rotated or inverted person object is corrected to an upright state in three stages to process the rotated scene pose estimation. Our experiments are tested in the COCO2017 dataset. The experimental results are shown in TABLE 1. Compared with other methods, the Average Precision (AP) of our method is higher than that of other methods, and the AP reaches 71.5%.

TABLE 1. Performance comparison of rotation-invariant module test on COCO test-dev dataset.

Method	AP
Mask-RCNN [10]	63.1%
G-RMI [35]	64.9%
RMPE [11]	68.8%
MultiPoseNet [36]	70.5%
PRCM (Ours)	**71.5%**

Visualization compare on rotation invariant is also shown in FIGURE 8. On the inverted person of image (a), Mask-RCNN incorrectly marks the key points of the face on the top of the image, but the top of the image is the position of the ankle rather than the face. Mask-RCNN [10] did not recognize this as an inverted person. For slanted athlete in image (c), neither Mask-RCNN nor RMPE [11] correctly marked the right hand of the athlete. The visualization comparison show that our results achieve good performance in the case of inverted and slanted characters in the image, as shown in the second column.

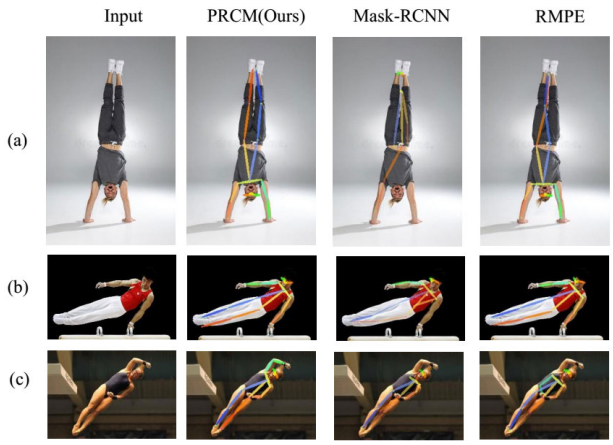


FIGURE 8. Visualization compare with state-of-the-art: The first column is the input image, the second column is the visualization of our method, and the third and fourth columns are the visualizations of Mask-RCNN and RMPE, respectively.

2) RESULTS ON SCALE-INVARIANT MODULE

To confirm the effectiveness of our proposed Scale-invariance module (SIM), we designed eight ablation experiments. ResNet50 and ResNet101 are adopted as our baseline.

In Experiment 1-4, ResNet50 was used as the backbone network. In Experiment 1, in the feature maps of different scales, all $d = 1$ dilated convolution modules were used, and the model can be expressed as (ResNet50 + d1). In Experiment 2, all the convolution modules with $d = 2$ are used, and the model can be expressed as (ResNet50 + d2). In Experiment 3, using the dilated convolution module with $d = 3$, the model can be expressed as (ResNet50 + d3). In Experiment 4, we use three different types of dilated convolutions simultaneously, and the model can be expressed as (ResNet50 + d1 + d2 + d3). Here, we will introduce the design scheme of Experiment 4 in detail. In the large-scale feature map C_1, first use the convolution kernel with $d = 3$, reduce the feature map to 1/2 of the original feature map, and then use the convolution kernel with $d = 2$ to reduce the original feature map to 1/4. Then, the convolution kernel with $d = 1$ is used to reduce it to 1/8 of the original feature map. Similarly, in the feature map C_2, the convolution kernel with $d = 2$ is used first, and then the convolution kernel with $d = 1$ is used; in the feature map C_3, the convolution kernel with $d = 1$ is directly used; the feature map C_4 remains unchanged.

After these feature maps pass through the above-mentioned dilated convolution, the size of the feature map at this time is 10×8, and the number of channels is 2048. At this time, the feature maps are fused. The fusion at this time is an add operation and does not increase the number of channels. Finally, three deconvolution operations are performed to generate a feature map with a size of 80×64 and a number of channels of 17. Similarly, Experiments 5-8 used ResNet101 as the backbone network. The experimental results are shown in TABLE 2.

From TABLE 2, we can see that the model size with ResNet50 as the baseline is 232M, and the model size with ResNet101 as the baseline is 320M. For different scale feature maps, different dilated convolutions are used. In ResNet50 and ResNet101, the highest Average Precision (AP) are 69.5% and 71.7% respectively, while both networks use different dilated convolutions. We can conclude from the table that it is better to use different dilated convolutions in different feature maps than to use only specific dilated convolutions. Because in large feature maps, the use of large dilated convolution means that the larger the receptive field, the better the ability to adapt to large objects. Conversely, in small feature maps, the receptive field should be appropriately reduced. What's more, the semantic information in different feature maps is different. For pose estimation, not all features are concentrated on the feature map of the last layer, and different key points can be distributed on feature maps of different proportions. Therefore, it is necessary to fuse the information of different feature maps to adapt to multi-scale input. So, the SIM designed by us can solve

TABLE 2. Results of different dilated convolution in ResNet50 and ResNet101 on COCO test-dev dataset.

Experiment	Models	AP	Model size
Experiment 1	ResNet50+d1	68.4%	232M
Experiment 2	ResNet50+d2	68.9%	232M
Experiment 3	ResNet50+d3	68.7%	232M
Experiment 4	ResNet50+d1+d2+d3	**69.5%**	232M
Experiment 5	ResNet101+d1	70.9%	320M
Experiment 6	ResNet101+d2	71.2%	320M
Experiment 7	ResNet101+d3	71.4%	320M
Experiment 8	ResNet101+d1+d2+d3	**71.7%**	320M

TABLE 3. Comparisons on COCO test-dev dataset.

Method	Backbone	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
CMU-Pose [7]	-	61.8%	84.9%	67.5%	57.1%	68.2%	66.5%
Mask-RCNN [10]	ResNet50-FPN	63.1%	87.3%	68.7%	57.8%	71.4%	-
G-RMI [35]	ResNet101	64.9%	85.5%	71.3%	62.3%	70.0%	69.7%
RMPE [11]	ResNet101	68.8%	87.5%	75.9%	64.4%	75.1%	73.6%
MultiPoseNet [36]	ResNet152	70.5%	87.7%	77.2%	66.1%	77.3%	74.9%
SimpleBaseline [37]	ResNet101	71.4%	89.3%	79.3%	68.1%	78.1%	**77.1**%
RIM + Ours	ResNet101	71.5%	90.2%	80.8%	69.5%	78.0%	76.8%
SIM + Ours	ResNet101	71.7%	90.2%	80.9%	69.8%	78.1%	76.8%
RIM + SIM + Ours	ResNet101	**72.0%**	**90.3%**	**81.1%**	**69.9%**	**78.1%**	77.0%

the multi-scale problem in the image. Compared with using the same convolution module alone, the multi-scale feature fusion module we designed is indeed effective.

Visualization of ablation experiments is also shown as FIGURE 9. Three different sizes of dilated convolution are used by ResNet101 + d1 + d2 + d3, compared to a single size of dilated convolution. There is a significant improvement in the performance of SIM using different dilated convolution, especially at some smaller, denser scales.

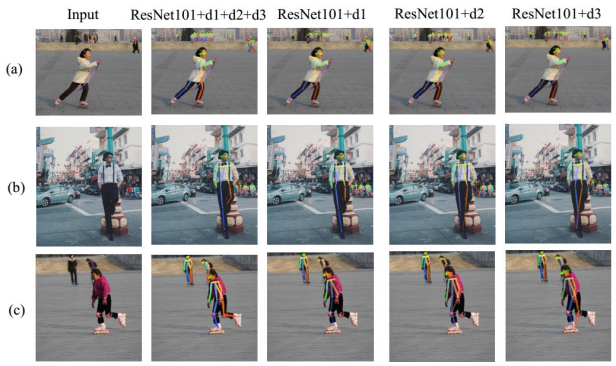


FIGURE 9. Visualization of ablation experiments: The first column is the input image and the second through fifth columns are the results of the ablation experiments, respectively.

3) COMPARISONS ON COCO2017 DATASET

To illustrate the effectiveness of our method, we have compared it with the state-of-the-art. For testing the proposed overall network, we designed three experiments. The first is to add only the Progressive rotation correction module (PRCM), which can be expressed as (PRCM + Ours). The second experiment is to add only the Scale-invariant module (SIM), which is the result of Experiment 8 above. This experiment can be expressed as (SIM + Ours). The third is to add both PRCM and SIM, which can be expressed as (PRCM + SIM + Ours).

For our baseline here, a human detector with person detection Average Precision (AP) of 56.4% on COCO std-dev split dataset is used. For reference, CPN [9] uses a human detector with person detection AP of 62.9% on COCO mini-val split dataset and SimpleBaseline [37] uses a human detector with person detection AP of 62.9%. Compared with CMU-Pose [7], Mask-RCNN [10] and MultiPoseNet [36], our method is significantly better. Compared to SimpleBaseline [37], their human detection is better, but overall AP is lower than us. Compared to MultiPoseNet [36], their feature extraction network is ResNet152, but the overall AP is still lower than ours. The results of other methods are summarized in TABLE 3 in the literature on the COCO2017 dataset.

We also tested on the MPII dataset and achieved state-of-the-art results across all key points on the MPII Human Pose dataset. Our network use ResNet101 as the backbone network

TABLE 4. PCKh@0.5 results Comparisons on MPII dataset.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Total
Wei et al. [6]	97.8%	95.0%	88.7%	84.0%	88.4%	82.8%	79.4%	88.5%
Newell et al. [5]	98.2%	96.3%	91.3%	87.1%	90.1%	87.4%	83.6%	90.9%
Yang et al. [38]	98.5%	96.7%	92.5%	88.7%	**91.1%**	88.6%	86.0%	92.0%
Ke et al. [39]	98.5%	**96.8%**	92.7%	88.4%	90.6%	**89.3%**	**86.3%**	92.1%
RIM + SIM + Ours	**98.6%**	**96.8%**	**92.8%**	**88.9%**	91.5%	89.1%	86.2%	**92.2%**

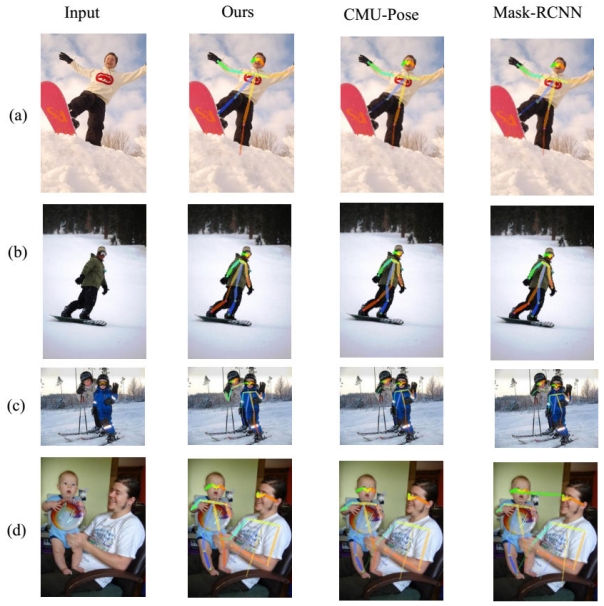


FIGURE 10. Visualization compare with state-of-the-art: The first column is the input image, the second column is the visualization of our method, and the third and fourth columns are the visualizations of CMU-Pose and Mask-RCNN, respectively. The input images (a) and (b) are rotated state, and (c) and (d) are different scales case.

and joins PRCM and SIM, which can be represented as (RIM + SIM + Ours). Evaluation is done using the standard Percentage of Correct Keypoints (PCK) metric which reports the percentage of detections that fall within a normalized distance of the ground truth. For MPII, distance is normalized by a fraction of the head size (referred to as PCKh [5]). Our results can be seen in TABLE 4, and the results on MPII are very competitive reaching 98.6% at PCK@0.5 accuracy on the head. The other methods do not deal with the issues caused by rotation and inversion. Instead, our method takes into account these two special cases. The experiments indicate that our proposed method is indeed effective and achieves good results.

The visualization comparison with state-of-the-art is shown in FIGURE 10. The figure below is our result. For example, in (a), the CMU-Pose [7] method does not detect well on the right foot. Besides, the CMU-Pose [7] method does not correctly detect the buttocks in (b). In (c), Mask-RCNN [10] method detects poorly the occluded left knee. In image (d), Mask-RCNN [10] cannot detect eyes well, and there is confusion. By contrast, we can see that our results (in (d)) are better than others with almost error-free. In summary, experimental results show that the proposed method is effective to tackle with the issues caused by rotation and large-scale difference.

V. CONCLUSION

In this paper, we have proposed a novel multi-person pose estimation under complex environment based on progressive rotation correction and multi-scale feature fusion. The pose estimation problem of rotating image and multi-scale person object in images are solved by our algorithm. First of all, we use a top-down approach, and then detect each person from the image. Second, the single person image is sent to a Progressive rotation correction module (PRCM), which solves the problem of rotating or inverted human image. The corrected image will be output after passing through the PRCM module. Then, the corrected image is sent to a Scale-invariant module (SIM) based on multi-scale feature fusion. In this module, dilated convolutions with different receptive fields are used to extract information in multiple-scale. At the same time, considering the different amounts of information carried in feature maps of different sizes in multi-scale images, we adopt a multi-scale feature fusion method to solve the problem of scale invariance. Then, we designed ablation experiments and compared them with state-of-the-art. A large number of experiments show that our proposed PRCM and SIM can effectively solve the case of rotation and multi-scale. Compared with the existing methods, the proposed PRCM can achieve better detection results in the case of rotated or inverted. In addition, the proposed SIM also performs better than existing methods in multi-scale images. Overall, the AP of our algorithm reached 72.0% in the COCO2017 dataset. However, our method also fails to detect when the person object is distorted or too small in the image.

In the future, we hope to apply the PRCM to the Optical Character Recognition (OCR) and the block alignment of Person re-identification. At the same time, we will improve the robustness of the PRCM to better adapt to the distorted images in pose estimation. In addition, we will try to fuse feature maps of four different scales to extract deeper features, and make predictions layer by layer to solve the detection of smaller objects.

REFERENCES

[1] 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.

[2] S. Alyammahi, H. Bhaskar, D. Ruta, and M. Al-Mualla, "People detection and articulated pose estimation framework for crowded scenes," *Knowl.-Based Syst.*, vol. 131, pp. 83–104, Sep. 2017.

[3] 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. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1799–1807.

[4] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1653–1660.

[5] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 483–499.

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

[7] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7291–7299.

[8] A. Newell, Z. Huang, and J. Deng, "Associative embedding: End-to-end learning for joint detection and grouping," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 2277–2287.

[9] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7103–7112.

[10] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2961–2969.

[11] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2334–2343.

[12] W. Li, Z. Wang, B. Yin, Q. Peng, Y. Du, T. Xiao, G. Yu, H. Lu, Y. Wei, and J. Sun, "Rethinking on multi-stage networks for human pose estimation," 2019, *arXiv:1901.00148*. [Online]. Available: http://arxiv.org/abs/1901.00148

[13] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[16] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[17] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.

[18] M. Jaderberg, K. Simonyan, and A. Zisserman, "Spatial transformer networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.

[19] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.

[20] G. Cheng, P. Zhou, and J. Han, "RIFD-CNN: Rotation-invariant and Fisher discriminative convolutional neural networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2884–2893.

[21] D. E. Worrall, S. J. Garbin, D. Turmukhambetov, and G. J. Brostow, "Harmonic networks: Deep translation and rotation equivariance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5028–5037.

[22] T. Cohen and M. Welling, "Group equivariant convolutional networks," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 2990–2999.

[23] X. Shi, S. Shan, M. Kan, S. Wu, and X. Chen, "Real-time rotation-invariant face detection with progressive calibration networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2295–2303.

[24] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[25] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767

[26] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*. [Online]. Available: http://arxiv.org/abs/1511.07122

[27] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," 2019, *arXiv:1902.09212*. [Online]. Available: http://arxiv.org/abs/1902.09212

[28] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[29] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection snip," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3578–3587.

[30] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: http://arxiv.org/abs/1706.05587

[31] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[32] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," 2016, *arXiv:1602.00134*. [Online]. Available: https://arxiv.org/abs/1602.00134

[33] Y. Li, Y. Chen, N. Wang, and Z. Zhang, "Scale-aware trident networks for object detection," 2019, *arXiv:1901.01892*. [Online]. Available: http://arxiv.org/abs/1901.01892

[34] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[35] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy, "Towards accurate multi-person pose estimation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4903–4911.

[36] M. Kocabas, S. Karagoz, and E. Akbas, "MultiPoseNet: Fast multi-person pose estimation using pose residual network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 417–433.

[37] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 466–481.

[38] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1281–1290.

[39] 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)*, 2018, pp. 713–728.

GUOHENG HUANG (Member, IEEE) received the B.Sc. degree in mathematics and applied mathematics and the M.Eng. degree in computer science degrees from South China Normal University, in 2008 and 2012, respectively, and the Ph.D. degree in software engineering from Macau University, in 2017. He is currently an Assistant Professor of computer science with the Guangdong University of Technology. He has hosted and undertaken number of national and provincial-level scientific research projects, including the Natural Science Foundation of China and National Key Research and Development Plan. As a key member of the Guangdong Key Laboratory of Cyber-Physical System, he has published many research papers. His research interests include computer vision, pattern recognition, and artificial intelligence. He is a CCF member.

XIAOPING CHEN received the bachelor's degree in engineering from Gannan Normal University. He is currently pursuing the master's degree in computer science with the Guangdong University of Technology. His main research fields are computer vision and human pose estimation.

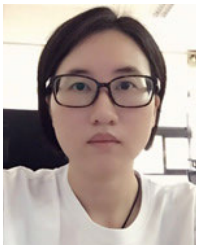
JUNAN CHEN is currently pursuing the bachelor's degree with the School of Computer Science, Guangdong University of Technology. He hosts an innovation and entrepreneurship project for college students. His main research field is computer vision.

WEIDA LIN received the bachelor's degree in engineering from the Binjiang College, Nanjing University of Information Science and Technology. He is currently pursuing the master's degree in computer science with the Guangdong University of Technology. His main research fields are computer vision and human pose estimation.

WING-KUEN LING (Senior Member, IEEE) received the B.Eng. (Hons.) and M.Phil. degrees from the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, in 1997 and 2000, respectively, and the Ph.D. degree from the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, in 2003. In 2004, he joined the King's College London as a Lecturer. In 2010, he joined the University of Lincoln as a Principal Lecturer and promoted to a Reader, in 2011. In 2012, he joined the Guangdong University of Technology as a Full Professor. He has published an undergraduate textbook, a research monograph, several book chapters, a book review published in an IEEE journal, more than 150 internationally leading journal papers, and more than 130 highly rated international conference papers. His research interests include time frequency analysis, optimization theory, symbolic dynamics, multimedia signal processing, and biomedical signal processing. He is a Fellow of IET, a China National Young Thousand-People-Plan Distinguished Professor, and a University Hundred-People-Plan Distinguished Professor. He was awarded the best reviewer prizes from the IEEE Instrumentation and Measurement Society, in 2008 and 2012. He serves in the technical committees of nonlinear circuits and systems group, digital signal processing group, and power electronics and systems group of the IEEE Circuits and Systems Community. He has also served as the Guest Editor-In-Chief of several special issues of highly rated international journals, such as *Circuits, Systems and Signal Processing* and *American Journal of Engineering and Applied Sciences*, and is currently an Associate Editor of *Circuits, Systems and Signal Processing, Journal of Franklin Institute, Measurement, IET Signal Processing, and Journal of Industrial Management and Optimization*.

CHI-MAN PUN (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in software engineering from the University of Macau, in 1995 and 1998, respectively, and the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, in 2002. He is currently an Associate Professor and the Head of the Department of Computer and Information Science, University of Macau. He has investigated several funded research projects and authored/coauthored more than 100 refereed scientific papers in international journals, books, and conference proceedings. His research interests include digital image processing, multimedia forensics and watermarking, pattern recognition, and computer vision. He is a professional member of the ACM. He has also served as the editorial member/referee for many international journals such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY.

LIANGLUN CHENG (Senior Member, IEEE) received the B.E. and M.S. degrees in automation from the Huazhong University of Science and Technology, Wuhan, China, in 1988 and 1992, respectively, and the Ph.D. degree in automation from the Chinese Academy of Sciences, in 1999. He is currently a Professor with the Guangdong University of Technology, a Computer Dean of the Guangdong University of Technology, a Doctoral Tutor, an excellent Teacher of Nanyue, and a national-level training target for the Thousand-Ten Thousand project of cross-century talents in Guangdong Province. He is an Executive Director of the Robotics Professional Committee of China Automation Association, a member of China Computer Federation, and a Vice Chairman of Guangdong Automation Association. His main research interests include knowledge graph, knowledge automation, and information physics fusion systems.



ZHUOWEI WANG received the Ph.D. degree in computer system architecture form Wuhan University, Wuhan, China, in 2012. She is currently an Associate Professor with the Institute of Computing, Guangdong University of Technology. Her research interests focus on high-performance computing, low-power optimization, and distributed systems.