Deep High-resolution Network with Double Attention Residual Blocks for Human Pose Estimation

Zhanqiang Hu1, Han Jin1, Yingxu Qiao1, and Fen Luo1

1School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo 454003 China

Corresponding author: Yingxu Qiao (e-mail: qiaoyingxu@hpu.edu.cn).

This work is supported by the Henan University Scientific and Technological Innovation Team Support Program (19IRTSTHN012) and by the Foundation of Henan scientific and technological project (192102210281).

ABSTRACT To improve the accuracy of human pose estimation, a novel method based on the deep high-resolution network (HRNet) and equipped with double attention residual blocks is proposed. Firstly, the channel attention and spatial attention modules are added to the residual block of feature extraction, resulting in the network paying more attention to the target area which needs to be extracted important information and suppressed unimportant information. Moreover, this paper proposes a novel module, Parallel Residual Attention Block (PRAB), which parallels the 3x3 group convolution of ResNeXt to the 3x3 convolution layer in the Bottleneck of ResNet, and then adds channel attention and spatial attention modules to these two branches respectively. In this way, the network can further improve the accuracy of human keypoint detection without significantly increasing the computation overhead. To demonstrate the effectiveness of our method, a series of comparative experiments are conducted on the MPII Human Pose dataset and the COCO2017 keypoint detection dataset. Experimental results illustrate that the attention mechanism is effective to improve the accuracy of human pose estimation and the proposed PRAB obtained the best results 90.5% on MPII which outperforms the existing methods.

INDEX TERMS Human pose estimation, attention mechanism, residual block, PRAB.

I. INTRODUCTION

Human pose estimation is a basic research topic in computer vision, which has a broad application in behavior recognition, human-computer interaction, automatic driving, etc. At present, the human pose estimation of the static image is to input an image into a trained network model, and then output the accurate pixel position of the human body key points (e.g. head, wrist, knee, ankle, etc.). Human pose estimation in the static images is the basis of video pose estimation and tracking [1-4], which means it is very useful for higher-level tasks such as action recognition [5].

In the past decade, the rapid development of deep learning has brought more possibilities to solve the problem of human pose estimation. Both the design of the network model [6-9] and the openness of the mainstream dataset [10-11] have made great contributions to the research of human pose estimation. Therefore, more and more powerful algorithms are constantly proposed, and the accuracy of human keypoint detection is constantly improved. However, due to the influence of occlusion, illumination, and complex background, etc., the detection of some hard key points of the human body is still not accurate enough.

Based on deep learning, there are two pipelines to estimate human pose: top-down [2, 6-8, 12] and bottom-up [13-15]. With top-down, the pedestrian detector is used to locate the pedestrian accurately, and then the key points of each pedestrian are detected separately, and bottom-up is to directly predict all the key points of all the people in the image and then according to some certain rules and strategies to allocate and connect the key points to get the human pose prediction. The algorithm of human pose estimation based on these two methods has achieved good results. Although the latter is faster, it is easy to miss some small and medium-sized human bodies’ detection in the image, while the former is more accurate in detecting human key points. Most of the existing top-down network models are based on high-resolution to low-resolution and then to high-resolution [2, 6-7], that is, to
recover high-resolution representation from low-resolution representation to estimate human pose. HRNet [12] has provided prior knowledge on how to maintain high-resolution characterization throughout the process for human pose estimation. It is recognized as the baseline model in pose estimation tasks since it has an outstanding capability.

Attention mechanism has been successfully applied in image processing, speech recognition, and natural language processing in recent years. The attention mechanism in computer vision is a special brain signal processing mechanism in human vision. Human vision can scan the global image to capture the target area that needs to be focused on, and then invest more attention resources in this area to obtain more details of the target that needs to be focused on and suppress other useless information. The attention mechanism greatly improves the efficiency and accuracy of visual information processing and provides the possibility to further improve the accuracy of human pose estimation.

There is a lot of redundant information in an image when estimating human pose. Therefore, the performance and efficiency of the network model will degrade if the feature extraction of irrelevant information is undifferentiated. In this paper, the top-down method is used to study how to design effective attention modules to further improve the accuracy of human keypoint detection. Here, we use HRNet [12] as the backbone of our model and employ its basic component residual block [16] for feature extraction. Then we add a dual attention module in each residual block, namely channel attention and spatial attention [17-19]. At the same time, a new parallel residual block PRAB (Group convolution operation [20] with a double attention module in parallel) is put into the start stage of the network. This enables the network model to parse out what information in the image is more important for the task at low level. And with the deepening of the network layer, more attention is focused on the important information while the secondary information can be suppressed. The network model equipped with channel attention and spatial attention mechanism can achieve the above objectives both in channel dimension and spatial dimension. Finally, the overall representation capability of the network model for human pose estimation task is improved.

The network structure is described in Figure 1 to illustrate the design of our feature extraction module. To demonstrate the effectiveness of our proposed method, experiments are conducted on two mainstream human pose estimation datasets MPII [10] and COCO2017 [11]. The experimental results revealed that the dual attention mechanism is effective to improve the accuracy of human pose estimation. Also, comparative experiments on different design proposals of attention mechanism proved the proposed PRAB was the most reasonable scheme with the best results.

The remainder of the paper is organized as follows. After the related work discussion in Section II, we cover the details of our proposed method in Section III. Section IV introduces experimental designs and discusses the results. We conclude with a short discussion in Section V.

**Figure 1.** Illustrating the architecture of our human pose estimation network. We replaced some residual blocks of HRNet with our proposed dual attention residual blocks and PRAB. See Section III D for details.
II. RELATED WORK

A. Human Pose Estimation

Traditional human pose estimation algorithms are mainly based on tree model and graph model to solve the problem of human key point detection [21], such as Random Field Models [22] and Dependency Graph Models [23]. These methods usually need to design some artificial features, which are simple, intuitive, and efficient. However, due to the considerable deformations of the human body and the complexity and diversity of the image itself, the accuracy of traditional methods is still unsatisfied, and cannot meet the demands of practical application.

Some researchers [24-26] discover that CNN’s shallower layers focus on low-level texture and spatial information which can help the model determine the location of the target. Until 2014, DeepPose [27] first used a convolutional neural network to solve the problem of human posture estimation and proposed a cascade approach for a more accurate posture estimator. Subsequently, more and more researches and proposals have been made on the human pose estimation based on deep learning.

The top-down classical algorithm CPM [6] adopts a multi-stage detection strategy. It proposed to use a neural network to learn image features and spatial information at the same time, which makes end-to-end learning possible. It also uses intermediate supervision to solve the problem of gradient vanishing, and adds supervision to the feature map at the end of each stage. Since then, intermediate supervision has become a standard technique in multistage networks. Hourglass [7], for example, captures and integrates information at all scales of an image through repeating bottom-up and top-down processes, which cooperates with supervision to improve network performance.

G-RMI [28] used the method of object detection to estimate the human pose. It first uses Fast-RCNN [29] to predict the position and size of the human body in the image and then estimates the keypoints contained in each human box. In 2018, the human keypoint detection framework based on CPN [8] structure proposes to use two stages of GlobalNet and RefineNet to detect key points and this method effectively alleviates the detection problem of hard key points. The network structure is divided into several stages, and the first stage produces the preliminary detection results of key points. The subsequent stages are all fed with the prediction output of the previous stage and the features extracted from the original image. The purpose of this method is to alleviate some challenges of keypoint detection, such as occlusion.

Simple Baseline [2] proposed two years ago is a very simple model, which can be used for multi-person pose estimation and tracking. The model employs the detection results of the previous frame to fuse into a new frame, thus reducing some error detection. It can achieve the performance as that of networks [6-7] by simple down-sampling and up-sampling. [4] can also be used for video pose estimation. It used 2D estimated skeleton points directly instead of an image to obtain information. This method significantly improves the performance of the dataset where subjects are in movement. Recently, [30] relies on a cascade of two models and applies a web-shaped model over the detected landmarks to associate each landmark with a specific target area. This method detects the target at a reasonable distance and resolution to capture the best frame in the video.

Most of the above methods obtain high-resolution output by performing down-sampling and up-sampling, while the network structure proposed by HRNet [12] maintains high-resolution representation by connecting multiple subnetworks in parallel. It enhances high-resolution representation by multi-scale fusion, which provides higher accuracy for human pose estimation.

In conclusion, these networks based on deep learning for human posture estimation have been greatly improved in accuracy and efficiency compared with the traditional methods. Although they have advantages in the network model, the network does not focus on the analysis of the importance of different parts of the image to the task performed, and how to treat the information with different importance. The attention mechanism is helpful to solve the above problems. We observed that HRNet [12] framework uses standard residual blocks for feature extraction, which cannot distinguish important information and secondary information in the image. Therefore, we propose a method of introducing a dual attention mechanism into HRNet [12] to further improve the accuracy of human posture estimation, and we study what kind of attention mechanism is useful and why it can achieve the optimal results.

B. Attention Mechanism

Attention mechanism has been widely used in computer vision in recent years since it can provide higher accuracy for image classification and object detection. The attention mechanism is mainly divided into channel attention focusing on different features of the image and spatial attention focusing on different areas of the image. The essence of the channel attention mechanism is to model the importance of each feature. For different tasks, feature allocation can be based on input, which is simple and effective.

The representative algorithm SENet [17] is to get the weights of each feature channel through the operation of Squeeze and Excitation, to enhance the importance and weaken the unimportant features, and to make the extracted features more directional. The spatial attention mechanism is to find the most important part of the image for processing. BAM [18] and CBAM [19] increase spatial attention based on channel attention and further improve the overall network performance. Therefore, adding channel attention and spatial attention to the human pose estimation algorithm can help the model give different weights to input features, extract more critical information, and improve the overall performance of...
the network. Recently, [31] proposed an adaptive-threshold-based multi-model fusion network, which combines the advantages of different deep learning methods to compress the object, to generate a high-quality image. [32] proposed a hierarchical residual block (HDB) for feature extraction and expression in efficient information and parameter-sharing fashion. All of these methods have improved the quality of feature representation.

Therefore, adding channel attention and spatial attention into the human pose estimation network can help the network model to give different weights to the features extracted from different parts of the image, and pay more attention to the information which is useful to the task. Most existing networks use attention mechanisms by adding them in series to a standard residual block or a layer of the network. In this paper, we try to characterize the feature difference of the image from the shallow layer, and with the deepening of the network layer, we continue to emphasize the important information and suppress unimportant information. We chose to add double attention mechanisms to each residual block of HRNet [12] while placing the parallel residual block PRAB at the start stage of the network. The experimental results show that the overall performance of the network is improved without significantly increasing the computational cost.

III. METHOD

HRNet [12] is an existing high-resolution network with the highest performance at present, which maintains high-resolution characterization in the whole process. It starts from a high-resolution subnetwork as the first stage, gradually add high-to-low resolution subnetworks one by one, forming new stages, and connect the multi-resolution subnetworks in parallel. At the same time, multi-scale fusion is repeated at the end of each stage, which makes each subnetwork repeatedly receive the information from previous parallel subnetworks. This kind of network design can continuously obtain rich high-resolution feature representation, and finally, obtain a more accurate heatmap of key points. Given the above advantages, we take HRNet [12] as the network framework for human pose estimation, and further improve the accuracy of human keypoint detection.

Although HRNet [12] dose feature fusion for multi-resolution subnets, the whole network treated all information of the image in feature extraction without discrimination. Due to not all the information in an image is equally important when detecting the keypoints, we should distinguish the information by increasing the weight of the more important information and reducing the interference of the secondary information. To extract the important content and location information of human keypoints, we add channel attention and spatial attention to all residual blocks. At the same time, PRAB (a parallel improvement on Bottleneck [16]) is proposed, and a convolution operation with two attention mechanisms is added [20]. This method is described in detail as follows:

A. Channel Attention Module

Channel attention (CA) is to generate channel attention maps by using the relationship between channels. Each channel of feature map is considered as a feature detector [33], so channel attention is used to solve the problem of ‘what’ is meaningful for a given input image.

In our method, firstly, we add channel attention to the residual blocks of HRNet [12] and rescale the features of each channel by modeling the interdependency between feature channels. This can make the network focus more on useful features.
Figure 4. The architecture of the main residual blocks. (a) A Basicblock of ResNet. (b) A Basicblock with double attention. (c) A Bottleneck of ResNet. (d) A Bottleneck with double attention. (e) A block of ResNeXt with cardinality=32. (f) PRAB which is proposed in our method.

channels and enhance the capability of distinguishing learning. As shown in Figure 2, we use average-pooling and max-pooling operations separately on the spatial dimension of the input feature map $F$ and generate two different spatial background descriptions: $F_{avg}$ and $F_{max}$.

\[
F_{avg} = \text{AvgPool}(F) \tag{1}
\]

\[
F_{max} = \text{MaxPool}(F) \tag{2}
\]

Then, we pass the two-channel descriptions through two convolution layers $fc_1$ and $fc_2$, which are parameter shared. The number of channels in the first layer is reduced to $C/r$, and that of the second layer is restored to $C$. The ReLU activation function is followed by $fc_1$. Then they are added up and passed through the sigmoid function, the weight coefficient $Mc(F)$ is finally obtained.

\[
Mc(F) = \sigma(fc_2(\text{ReLU}(fc_1(F_{avg}))))
+ fc_2(\text{ReLU}(fc_1(F_{max})))) \tag{3}
\]

where $fc_1$ and $fc_2$ use $1 \times 1$ convolution operation to reduce and expand the feature channel. And $r$ is the scaling ratio, $\sigma$ denotes the sigmoid function. Finally, we multiply the weight coefficient $Mc(F)$ by the original feature $F$ to get the channel attention map $Fc$.

\[
Fc = Mc(F) \times F \tag{4}
\]

B. Spatial Attention Module

Spatial attention (SA) is to generate spatial attention maps by using the inter-spatial relationship of features. Because the contribution of each area is not equal in the input image, only the area related to the task needs to be concerned, so spatial attention focuses on the important places of the input image.

In our method, after the channel attention, spatial attention is added, which improves the accuracy of extracting image features from two aspects of ‘what’ and ‘where’. We use the average-pooling and max-pooling operations separately along the channel axis of feature map $Fc$, and generate two descriptions: $F_{avg}$ and $F_{max}$.

\[
F_{avg} = \text{AvgPool}(Fc) \tag{5}
\]

\[
F_{max} = \text{MaxPool}(Fc) \tag{6}
\]

Then the two maps are concatenated according to the channel. Then, through a $7 \times 7$ convolution layer, the sigmoid activation function, and the weight coefficient $Ms(Fc)$ is obtained, as shown in Figure 3.

\[
Ms(Fc) = \sigma(f^{7 \times 7}(\text{Concat}([F_{avg}, F_{max}]))) \tag{7}
\]

Finally, the spatial attention map $Fcs$ can be obtained by multiplying the weight coefficient $Ms(Fc)$ with the $Fc$.

\[
Fcs = Ms(Fc) \times Fc \tag{8}
\]

C. Double Attention Residual Block

Through the above analysis, we know that if we want to obtain more useful information for the human keypoint detection, we should not only analyze the content of the image but also analyze the usefulness from the relationship between different positions in the image. Therefore, this paper uses the residual block of the dual attention mechanism as the basic block of HRNet [12] to extract features. Channel attention and spatial attention were serially equipped to Basicblock and Bottleneck [16]. The detailed operation is shown in Figure 4.

A Basicblock of ResNet [16] is shown in Figure 4(a), which has only two $3 \times 3$ convolution layers. We add channel attention and spatial attention modules followed by the second convolution layer, as shown in Figure 4(b). A Bottleneck of ResNet [16] is shown in Figure 4(c), which contains two $1 \times 1$ convolution layers and one $3 \times 3$ convolution layer. And then we add channel attention and spatial attention modules between the $3 \times 3$ convolution and the second $1 \times 1$ convolution, as shown in Figure 4(d). In this way, the improved double attention residual block can pay more attention to the key areas of the image, suppress the useless information, and increase the accuracy of feature extraction.
D. Parallel Residual Attention Block

Analyzing the experimental results of adding the above-mentioned double attention residual block in HRNet [12] leads to a question. Is there a way to continuously improve the overall performance of the model without increasing the computational cost? The group convolution operation proposed by ResNeXt [20] network is helpful to solve this problem. The advantage of group convolution is to improve the accuracy of the model without increasing the complexity of parameters and reduce the number of hyperparameters. Therefore, a 3×3 group convolution is paralleled with the standard 3×3 convolution operation in Bottleneck [16], and a double attention module is added after the two convolutions. A detailed description of PRAB is given below.

The residual block for ResNeXt [20] is shown in Figure 4(e). In our method, we make parallel improvements to the Bottleneck as shown in Figure 4(f). It means that the Bottleneck is divided into two branches after the first 1×1 convolution layer, one of which still performs the 3×3 convolution operation of the Bottleneck of ResNet [16]. The other branch is the 3×3 group convolution operation [20]. Then, the channel attention module and spatial attention module are respectively connected after the two branches. Finally, we add the two branch outputs and through the second 1×1 convolution layer, so that we can recover the initial channel number, and then add the original input value to get the final value. We call the final residual block PRAB (Parallel Residual Attention Block). We put PRBA only at the start stage of the HRNet [12] network. It not only does not significantly increase the computational burden but also further improves the accuracy of the model in detecting key points of human body. We compared it with other ways of adding attention. For example, only one attention mechanism is added, or only parallel convolution operation is performed instead of two attention modules. Experimental results show that PRAB is the most effective.

E. Details of the Human Pose Estimation Network

The network framework of this paper consists of five stages, as shown in Figure 1. In the first stage, the height (H) and width (W) of the input image are changed to H/4 and W/4 by two convolution operations, the number of channels is 64. Then we use four double attention residual blocks (as shown in Figure 4(d)) or four PRAB (as shown in Figure 4(f)) to extract features, and the number of channels becomes 256. In the second stage, the channel number of the feature map is changed to 32 by convolution operation along the high-resolution subnetwork line. A branch with low resolution of 64 channels is generated based on the previous stage. Then the two branches use four double attention residual blocks (as shown in Figure 4(b)) to extract features, and multi-scale fusion is adopted at the end of this stage. This means that the final output of the low-resolution branch is obtained by down-sampling the high-resolution branch and adding it to the low-resolution. At the same time, the low-resolution branch is up-sampled to the high-resolution branch through the simple nearest-neighbor sampling, then adding it to the high-resolution branch to get the final output of the high-resolution branch. The third and fourth stages are the same as the second stage, both of which generate new low-resolution branches, and the number of channels is 128 and 256 respectively. Four branches can be obtained after feature extraction and a multi-scale input map is fed into the fifth stage. In the fifth stage, the three low-resolution branches are all up-sampled and added to the high-resolution branch with 32 channels. Finally, we adopt a 1×1 convolution to get the ultimate output of the network model. The result is a keypoint heatmap, and the final number of channels is the number of keypoints in the human pose estimation dataset.

IV. Experiments

Our method is conducted on the MPII Human Pose dataset [10] and the COCO keypoint detection dataset [11], which demonstrates that it can achieve better accuracy. In the experiments, we compare our human pose estimation network based on HRNet [12] with double attention residual blocks and other models. Besides, we also compared different ways of adding attention mechanisms and different parallel operations. Finally, to verify the effectiveness of the proposed PRAB, we replaced the residual block of Simple Baseline [2] with PRAB as a complementary experiment. All the above experimental results prove the effectiveness of PRAB. The detailed experimental design and experimental results are as follows.

A. The experimental details

To make a fair comparison, all experiments in this paper are implemented in the same experimental environment. We use Pytorch to build the program code and run the model in a server with two NVIDIA 2080ti GPU, and an i9-9900k CPU with 8 cores. The model uses the ImageNet pre-trained model. We use Adam optimizer with a batch size of 16 to update the parameters. Batch normalization is also used to improve training. The initial learning rate of the model training is set as 1e-3 and is dropped to 1e-4 and 1e-5 at the 170th and the 200th epoch respectively on the MPII dataset. The training finally stops at the 310th epoch. On the COCO dataset, the initial learning rate of the model training is also set as 1e-3 and is dropped to 1e-4, 1e-5, 1e-6 at the 170th, 220th, 230th epoch respectively, and the training finally stops at the 250th epoch.

The proposed network uses a Mean Squared Error (MSE) as the loss function to compare the predicted heatmap with the ground-truth heatmap. In the training process of the model, the data augmentation of horizontal flipping, random scaling, and random rotation is used. Each keypoint location is predicted by adjusting the highest heat value location with a quarter offset in the direction from the highest response to the second highest response.
B. Experiments on MPII Dataset

The MPII Human Pose dataset contains 25K images in which 40K human instances are labeled with 16 key points. We divided the whole dataset into a 22K training set and a 2975 test set in our experiments. We fixed the aspect ratio of the provided person boxes to 4:3, then crop the box from the image and resized it to a fixed size, 256 $\times$ 256. The data augmentation includes a random rotation in (-45°, 45°), random scaling in (0.65, 1.35), and horizontal flipping. The PCKh score based on the normalization of the head size is used to measure the accuracy of the model in the MPII dataset.

1) COMPARE WITH THE STATE-OF-THE-ART AND CLASSICAL ALGORITHMS.

As shown in Table I, we compare our two proposed methods with HRNet [12] and the classical human pose estimation algorithms developed in the past years. The experimental results show that our method is the best in keypoint detection. The two improved methods are: (i) adding two attention modules (CA + SA) directly to Basicblock and Bottleneck; (ii) putting the PRAB in the start stage of the network instead of Bottleneck. The details are described in part III.D and III.E. We provide the PCKh@0.5 results in Table I. The overall results of our two methods on the test set are 90.4% and 90.5%, which are better than HRNet [12] 0.1% and 0.2% respectively. It can be proved that the dual attention mechanism is effective for human pose estimation. Besides, the values in Params and GFLOPs columns reflect that our method has a slight increase in the number of parameters, but the model computational complexity remains unchanged compared with HRNet [12]. This demonstrates that our double attention module can further improve the performance of the network without significantly increasing the computational cost.

We compare the accuracy of our (PRAB) and HRNet [12] in the middle process of model training. As can be seen in Figure 5, the accuracy of our model is better than HRNet [12] after the 170th epoch. It can also be found that the network model is more stable in the whole training process. We also compare the mean values of all detected key points in the whole training process of the proposed model with HRNet [12], shown in Figure 6. Similarly, in the later process of model training, our method is better than HRNet [12].

![Figure 5. Comparison results of accuracy between our(PRAB) and HRNet.](image1)

![Figure 6. The comparison of the mean value of detected keypoints between our (PRAB) and HRNet in the whole iteration process.](image2)
Figure 7. The visual comparison of our (PRAB) and HRNet on the MPII dataset.

TABLE II

| Method            | Head | Sho | Elb | Wri | Hip | Knee | Ank | Mean |
|-------------------|------|-----|-----|-----|-----|------|-----|------|
| ResNet(CA)        | 97.2 | 96.1| 90.6| 86.6| 89.4| 86.2 | 82.9| 90.3 |
| ResNet(SA)        | 97.2 | 96.0| 90.5| 86.4| 89.0| 86.5 | 83.0| 90.3 |
| ResNeXt(CA+SA)    | 97.1 | 95.9| 90.6| 86.3| 89.3| 86.3 | 82.8| 90.2 |
| ResNet'(CA+SA)    | 97.0 | 95.9| 90.7| 86.2| 89.6| 86.5 | 82.6| 90.2 |
| Our(CA+SA)*       | 97.1 | 96.0| 90.7| 86.4| 89.4| 86.8 | 83.4| **90.4** |
| PRAB(CA)          | 97.2 | 95.7| 90.4| 85.9| 89.3| 86.4 | 82.9| 90.1 |
| PRAB(SA)          | 97.0 | 96.1| 90.7| 86.2| 89.0| 86.7 | 83.0| 90.2 |
| PRAB'(CA+SA)      | 97.0 | 96.1| 90.7| 85.9| 89.0| 86.9 | 83.2| 90.3 |
| Our(PRAB)*        | 97.1 | 96.2| 90.7| 86.4| 89.8| 86.9 | 83.3| **90.5** |

2) COMPARISON OF DIFFERENT WAYS OF ADDING ATTENTION MECHANISMS AND DIFFERENT PARALLEL OPERATIONS.

The experimental results in Table II further demonstrate the effectiveness of the double attention mechanism and our proposed PRAB. (i) Compared with only adding channel attention ResNet (CA) and spatial attention ResNet (SA) separately, our method Our (CA+SA) performs better on the test set, which because our method makes the network pay more attention to useful information from two dimensions of channel and space. This method is also better than adding double attention modules to the block of ResNeXt [20] (ResNeXt (CA + SA)). Furthermore, double attention modules are added to Bottleneck's 3x3 convolution method (ResNet (CA + SA)) is not as good as Ours (CA+SA) too. (ii) In the same way, Our (PRAB) also performs better than the parallel improved residual blocks with only adding channel attention...
### Table III

Experiment results with different values of $r$ and the value are PCKh@0.5 scores on MPII. All values in the table are in percentages.

| Value | Head | Sho | Elb | Wri | Hip | Knee | Ank | Mean |
|-------|------|-----|-----|-----|-----|------|-----|------|
| $r=8$ | 97.4 | 96.0 | 90.3 | 85.7 | 88.8 | 86.7 | 83.0 | 90.1 |
| $r=16$ | 97.1 | 96.2 | 90.7 | 86.4 | 89.8 | 86.9 | 83.3 | 90.5 |
| $r=32$ | 97.1 | 96.1 | 90.7 | 86.3 | 89.1 | 86.7 | 82.9 | 90.3 |

### Table IV

Experiment results of different human pose estimation method on the COCO val2017.

| Method          | Params | GFLOPs | AP  | AP.5 | AP.75 | AP<sub>5</sub> | AP<sub>L</sub> | AR  |
|-----------------|--------|--------|-----|------|-------|---------------|---------------|-----|
| SimpleBaseline-152[2] | 68.6M  | 15.7   | 72.0 | 89.3 | 79.8  | 68.7          | 78.9          | 77.8|
| HRNet-W32[10]   | 28.5M  | 7.1    | 74.4 | 90.5 | 81.9  | 70.8          | 81.0          | 79.8|
| Our(PRAB)       | 28.9M  | 7.1    | **74.6** | 90.7 | **82.1** | **71.1**      | **81.3**      | **80.0**|

* AP.5, AP.75 (AP at OKS=0.5, 0.75), AP<sub>5</sub> for medium objects, AP<sub>L</sub> for large objects.

**Figure 8.** Visualization results of human pose estimation for some images.

(PRAB(CA)) and spatial attention (PRAB (SA)). PRAB(CA+SA) in Table II is to parallel the 3×3 convolution operation in ResNeXt[20] with the 3×3 convolution operation in Bottleneck, sum them up and pass the 1×1 convolution the operation, then adds two attention modules. This method only adds one more double attention block. The experimental results also show that it is not as good as Our (PRAB), which is parallel two branch double attention modules. In conclusion, our proposed method based on PRAB achieves the best results. To sum up, the comparison experiments by adding different attention mechanisms and different parallel operations, illustrate that the proposed method based on PRAB is the most effective network and can achieve the best performance.

3) **Experiments on Different Values of $r$.**

We have done a series of experiments on different values of the scaling ratio $r$ in the channel attention module of our network (Our (PRAB)). The experimental results in Table III show that the best results are obtained when $r$ is set to 16.

C. **Experiments on COCO Dataset**

COCO2107 key point detection dataset contains 200K images in which 250K person instances are labeled with 17 key points. Here, the train2017 set includes 57K images and 150K person instances, and the val2017 set has 5K images. The data augmentation is the same as MPII, except the person boxes are cropped to 256×192. The standard evaluation metric OKS (Object Keypoint Similarity) is used to measure the similarity.
between the predicted key points and the real key points, then we report the average precision (AP) and recall scores (AR).

The results in Table IV show that our network based on PRAB performs better than HRNet [12] and Simple Baseline [2] on COCO val2017. The AP value and AR value of our method are 74.6% and 80.0% respectively, both of which are 0.2% higher than HRNet [12]. From the Params and GFLOPs columns in Table IV, it can be seen that our method further improves the network performance without significantly increasing the computational cost. Figure 8 describes the visual detection results on the images in the COCO dataset.

| Method       | Head | Sho | Elb | Wri | Hip | Knee | Ank | Mean |
|--------------|------|-----|-----|-----|-----|------|-----|------|
| SimpleBaseline-50[2] | 96.3 | 95.0 | 88.3 | 82.4 | 87.6 | 81.8 | 77.8 | 87.7 |
| Our(PRAB)    | 96.4 | 95.1 | 88.1 | 82.3 | 87.6 | 83.6 | 79.7 | 89.1 |

D. Additional Experiments
To verify the effectiveness of the proposed PRAB, we add the PRAB into Simple Baseline [2] and conducted a comparative experiment on the MPII dataset. Simple Baseline [2] is a simple structure network based on ResNet [16]. It is mainly through simple down-sampling and up-sampling to estimate human pose. The experimental results in Table V show that our PRAB is also more effective than this network structure. The overall detection performance is 0.4% higher than that of Simple Baseline [2].

V. CONCLUSION
In this work, we propose a novel human pose estimation method based on HRNet [12] and equipped with double attention residual blocks, which is aimed to further improve the accuracy of human pose estimation. On one hand, our network preserves the high-resolution advantage of HRNet [12]. On the other hand, our network pays more attention to the key areas of the input image and extracts more important information by adding channel attention and spatial attention modules to the blocks of feature extraction. Furthermore, PRAB proposed in this paper can further improve the accuracy of human keypoint detection without significantly increasing the calculation overhead. The experimental results demonstrate that the employment of a double attention mechanism can accurately locate the key points, and our method PRAB achieves the best results.

In the future, we plan to further optimize the network structure and improve the convergence speed since the current network is a little too complex and has too many parameters.

REFERENCES
[1] T. Pfister, J. Charles, and A. Zisserman, “Flowing Convnets for Human Pose Estimation in Videos,” in IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 1913-1921.
[2] B. Xiao, H. Wu, and Y. Wei, “Simple Baselines for Human Pose Estimation and Tracking,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Munich, Germany, 2018, pp. 472–487.
[3] G. Liu, S. Liu, K. Muhammad, A. K. Sangaiah, and F. Doctor, “Object tracking in varying lighting conditions for fog based intelligent surveillance of public spaces,” IEEE Access, vol. 6, pp. 29283-29296, 2018.
[4] P. Barra, C. Bisogni, M. Nappi, et al., “Gender classification on 2D human skeleton,” in 2019 3rd International Conference on Bio-engineering for Smart Technologies (BioSMART), Paris, France, 2019, pp. 1-4.
[5] C. Feichtenhofer, A. Pinz, and A. Zisserman, “Convolutionsal Two-Stream Network Fusion for Video Action Recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, Nev, USA, 2016, pp. 1933-1941.
[6] S. E. Wei, V. Ramakrishna, T. Kanade, et al., “Convolutional Pose Machines,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, Nev, USA, 2016, pp. 4724-4732.
[7] A. Newell, K. Yang, J. Deng, “Stacked Hourglass Networks for Human Pose Estimation,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Amsterdam, The Netherlands, 2016, pp. 483-499.
[8] Y. Chen, Z. Wang, Y. Peng, et al., “Cascaded Pyramid Network for Multi-Person Pose Estimation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Salt Lake City, UT, 2018, pp. 7103-7112.
[9] Li W, Wang Z, Yin B, et al., “Rethinking on Multi-Stage Networks for Human Pose Estimation,” 2019, arXiv:1901.00148. [Online]. Available: https://arxiv.org/abs/1901.00148v4, pp. 1-10.
[10] M. Andriluka, L. Pishchulin, P. V. Gehler, and B. Schiele, “2D Human Pose Estimation: New Benchmark and State of the Art Analysis,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Columbus, OH, USA, 2014, pp. 3686-3693.
[11] T. Lin, M. Maire, S. J. Belongie, et al., “Microsoft COCO: Common Objects in Context,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Zurich, Switzerland, 2014, pp. 740-755.
[12] K. Sun, B. Xiao, D. Liu, and J. Wang, “Deep High-Resolution Representation Learning for Human Pose Estimation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). Long Beach, CA, USA, 2019, pp. 5866-5869.
[13] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh, “Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, 2017, pp. 1302-1310.
[14] Alejandro Newell, Zhiho Huang, and Jia Deng, “Associative Embedding: End-to-End Learning for Joint Detection and Grouping,” 2017, arXiv:1611.05424. [Online]. Available: https://arxiv.org/abs/1611.05424, pp. 1-11.
[15] B. Cheng, B. Xiao, J. Wang, et al., “HighResHRNet: Scale-Aware Representation Learning for Bottom-Up Human Pose Estimation,” Presented at CVPR, 2020, [Online]. Available: https://arxiv.org/abs/1908.10357v3, pp. 1-10.
[16] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, Nev, USA, 2016, pp. 770-778.
[17] J. Hu, L. Shen, and G. Sun, et al., “Squeeze-and-Excitation Networks,” IEEE Trans Pattern Anal Mach Intell, vol. 42, no. 8, pp. 2011-2023, Aug. 2020, 10.1109/TPAMI.2019.2913372.
[18] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, “BAM: Bottleneck Attention Module,” 2018, arXiv:1807.06514. [Online]. Available: https://arxiv.org/abs/1807.06514, pp. 1-14.
[19] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, “CBAM: Convolutional Block Attention Module,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Munich, Germany, 2018, pp. 3-19.
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3044885, IEEE Access