Human pose estimation based on Improved High Resolution Network

Ying Bao¹, Manlin Zhang² and Xiaoming Guo ¹*

¹ Liaoning Petrochemical University, Fushun, Liaoning, 113000, China
*Corresponding author’s e-mail: panbin@lnpu.edu.cn

Abstract. Human pose estimation has become a hot problem in human-computer interaction and other intelligent application technologies, but the results obtained by previous deep learning networks are not very satisfactory when facing small-scale human instances. Therefore, in order to solve the problem of scale variation in human pose estimation, especially to pinpoint the keypoints of small-scale human instances, an improved High-Resolution Network (Improved HRNet) is proposed in this paper. The main improvement work in this paper is as follows: In this paper, a double attention mechanism is added in the forward transmission of the parallel sub-network, with the aim of assigning weights to the propagated information without changing the number of channels, assigning the information with high weights as useful information and reducing the interference caused by irrelevant information. In this paper, the network structure is validated using COCO dataset, and the average accuracy (Average Precision, AP) of Improved HRNet is 66.4, which is 2.3 higher than the average accuracy of High-Resolution Network (HRNet) with only 0.4% increase in parameters.

1. Introduction
With the gradual development in the field of computer vision, human pose estimation, as an indispensable component of human-computer interaction, has gradually entered the limelight. Human pose estimation tasks play different roles in different fields and are widely used for human analysis in intelligent surveillance, medical rehabilitation and sports. In human pose estimation, estimation of 2D human pose from a picture or a video is a very important fundamental stage, for example, human tracking, motion perception, 3D pose research or human-computer interaction applications need accurate 2D human pose estimation as a support. Therefore, the study of 2D pose estimation becomes particularly important.

Human pose estimation is performed in two steps, namely human target detection and human key detection. Currently, human pose estimation is mainly performed by deep learning methods, while the network used for human keypoints detection is mainly a deep convolutional neural network, and the methods of deep convolutional neural network can be subdivided into two types, the first one is the estimation of keypoints heatmap method and the second one is the regression of keypoints position method. The keypoints of the human instance can be learned by extracting the corresponding feature information after learning each image and then selecting the bit with the highest heat value as the keypoints.

In 2016 Stacked Hourglass Network (SHN)¹ was proposed, which consists of subnetworks in series according to a resolution from high to low. Thus, the stacked hourglass network is operated by downsampling and upsampling from high to low and then from low to high for different resolutions. And this process cannot fully utilize the spatial feature information effectively, resulting in partial loss...
of spatial feature information thus making the output high-resolution representations less than perfect. The cascaded pyramid network (CPN) proposed in the literature\cite{2}, compensates this drawback of SHN and is able to fuse the low and high resolution feature map information when using the operation of upsampling.

In 2019 Sun et al. proposed the high-resolution network (HRN)\cite{3}, HRN abandon the series connection method used in conventional networks in the past, and it uses parallel connection to connect different resolution subnetworks in parallel from high to low, which can effectively utilize feature information when realizing multi-scale fusion.

2. HRNet
The HRNet network model is a completely new architecture compared to previous network models. It always maintains a high-resolution feature sub-network throughout the network delivery. In order to realize the transmission of high-resolution features, it divides the network into multiple sub-networks, with the backbone being the high-resolution sub-network, which always maintains the same resolution as the input feature map, and then gradually adds low-resolution sub-networks through down-sampling operations, so that more stage sub-networks are formed to achieve the state of multi-stage sub-networks in parallel. The parallel low-resolution sub-networks continuously exchange information through fusion to achieve the information of the feature map in the high-resolution sub-network contains deeper subnetwork features, while at the same time the backbone subnetwork is always the high-resolution feature map transmission, so that less information loss, it achieves the effect of better estimation of human keypoints. The HRNet network is shown in Figure 1, where the horizontal and vertical directions correspond to the depth of the network and the scale of the feature map, respectively.

3. Improved HRNet
Inspired by SENet\cite{4}, this paper adds a channel attention mechanism to the forward passing process of parallel sub-networks. Channel attention is to calculate weights for each feature map so that the network focuses on the features with more useful information, reducing the interference brought by irrelevant information to the network for extracting features as well as exploiting the interdependence between features. The structure of channel attention is shown in Figure 2.
As shown in Figure 2, it is assumed that the input characteristic graph is $X$ and the size is $H' \times W' \times C'$. After the convolution operation, the size of the feature graph is $H \times W \times C$, where $F_{tr}$ represents the convolution operation. Then, a feature vector of $1 \times 1 \times C$ is obtained by global average pooling. The calculation process is shown in Equation (1): where the $F_{sq}(\cdot)$ is used to represent the global average pooling operation and to represent the importance degree of different channel features.

$$Z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j)$$

(1)

$U_c(i, j)$ represents the pixel value of $(i, j)$, where the location of the $c$ channel in the feature graph $U$. Then, the correlation between channels is modeled through two full connection layers to learn the degree of dependence on each channel and adjust the feature graph according to the different degree of dependence. The calculation process is shown in Equation (2).

$$s = F_{ex}(Z_c, W) = Sigmoid(W_2 ReLu(W_1 Z_c))$$

(2)

Where, $F_{ex}$ represents the sigmoid loss function, $W_1$ and $W_2$ are the weights of the two convolutional layers respectively, and their sizes are constantly adjusted through training and learning. Finally, feature graphs with different weights on different channels are fused through $F_{scale}$, and the calculation process is shown in Equation (3).

$$x_c = F_{scale}(u_c, s_c) = s_c \cdot u_c$$

(3)

Adding channel attention makes it possible to exchange information without reducing the dimension of the subnetwork. This method not only ensures the computational performance but also reduces the complexity of the network model. Adding channel attention mechanism can not change the size and number of channels of the feature graph, but multiply all pixels in the same channel by the same value. If a channel is important, then the weight of all elements in it is multiplied by a high coefficient, and vice versa.

In addition to channel attention mechanism, spatial attention mechanism is also added to optimize feature extraction. In this paper, spatial attention feature map is generated by using spatial relations among features. Different from the focus of channel attention mechanism, spatial attention focuses on the information of "the position on the feature map", and uses spatial attention to complement channel attention. To calculate spatial attention, this paper first carries out average pooling and maximum pooling along the channel direction, and fuses them to generate effective feature weights (the structure of spatial attention mechanism is shown in Figure 3. According to [5], pooling along the channel direction can highlight useful information in the feature.

![Spatial Attention Mechanism Flowchart](image)

**Figure 3** The spatial attention mechanism flowchart

The operation is as follows: we use the average pooling operation to obtain the feature graph $F_{avg} \in \mathbb{R}^{1 \times H \times W}$ with the number of channels of 1, where each pixel represents the average value of the...
1×1 "large and small feature graph of each channel direction, and use the maximum pooling operation to obtain the feature graph \( F_{\text{max}} \in \mathbb{R}^{3 \times H \times W} \), where each pixel represents the maximum value of the 1×1 large and small feature graph of each channel direction. Then, two pooled feature graphs are fused through a layer of conventional convolution operation to obtain the final spatial attention feature map \( M_s(F) \in \mathbb{R}^{H \times W} \), which is used to process the features processed by channel attention mechanism again to improve the feature performance at the spatial scale. The process of spatial attention mechanism is expressed as follows:

\[
M_s(F) = \sigma(f^{7 \times 7}([\text{AvgPool}(F)];[\text{MaxPool}(F)])) = \sigma(f^{7 \times 7}(F_{\text{avg}};F_{\text{max}}))
\]

(4)

Where, \( \sigma \) represents the Sigmoid function, and \( f^{7 \times 7} \) is 7 × 7 denoise convolution operation. The convolution operation of 7 × 7 was used to fuse the output feature graphs of the two pooled layers.

In this paper, dual attention mechanism module is added in the forward transmission process of parallel subnetwork, which can make the network learn channel importance and space importance respectively. In the process of feature transmission, the network will focus on the features with more useful information, reduce the interference brought by irrelevant information to the network feature extraction and the interdependence between features, and improve the accuracy of keypoints information prediction in the feature map. The structure of dual attention modules is shown in Figure 4.

\[
M' = \left( (M \times F_{\text{scale}}) \times M_{sf} \right)
\]

(5)

Where \( M' \) is the output feature of the dual attention mechanism module, \( M \) is the input of the dual attention mechanism, \( F_{\text{scale}} \) is the weight of each channel output by the channel attention module, and \( M_{sf} \) is the feature map output by the spatial attention module.

![Figure 4 Structure diagram of convolutional block attention module](image)

4. Experimental results and analysis

4.1. Experimental setup

Computer configuration: system version: Ubuntu 18.04; Processor: Intel Xeon (R) Platinum 8160 CPU * 4; Memory for installation: 250G. This code is developed using python 3.6 and requires NVIDIA GPU. This code was developed and tested with 4 NVIDIA P100 GPU cards.

4.2. Data sets

This article uses the COCO dataset. The COCO dataset contains more than 200,000 images and more than 250,000 human instance images labeled with 17 keypoints. This paper is a model training on COCO2017 training set, which includes more than 57,000 images and 150,000 images of human body instances. In addition, the network structure of this paper is evaluated on COCO2017 validation set and COCO2017 test set, which contains 5,000 images and 20,000 images respectively.

4.3. Image preprocessing

As the original images collected in the MSCOCO2017 dataset are of different sizes, image preprocessing and then training are required. Therefore, the pretreatment process is divided into the following two
parts: (1) The image size is cut from the data set image centering on the main human hip. The size of the image is cut to 256×192 and adjusted to a fixed ratio, and the height: width is 4:3, which is convenient for network training. (2) Data enhancement with random rotation (−45°, 45°) and random scaling scale (0.65, 1.35) was used to process the data. The data set image is clipped with the main human hip as the center. According to the ratio of height to width of 4:3, the size of the image is clipped to 384×288, so as to achieve the input effect required by network training.

4.4 Evaluation criteria
In this paper, experiments are carried out on COCO2017 data set, and there are two evaluation methods: (1) The key similarity (OKS) given by MS COCO is used for evaluation. (2) using the percentage of correct key points (PCK) as the experimental evaluation index.

4.5 Experimental verification and analysis
Table 1 summarizes the results of COCO2017 test development data set. As can be seen from the table, this paper uses Improved HRNet to surpass the previous methods and become the best bottom-up method (AP70.5). The performance of the improved high-resolution network proposed in this paper is better than that of the high-resolution network, and the parameters are only slightly increased (+0.4%).

| methods       | model     | input parameter | FLOPS | AP   | AP50  | AP75  | APM  | APL  |
|---------------|-----------|-----------------|-------|------|-------|-------|------|------|
| Multiscale testing | Hourglass       | 512              | 277.8M | 206.9 | 63.0  | 85.7  | 68.9  | 58.0  | 70.4  |
| PersonLab     | Hourglass       | 512              | 68.7M  | 405.5 | 68.7  | 89.0  | 75.4  | 64.1  | 75.5  |
| Improved HRNet | HRNet-W32       | 512              | 28.6M  | 47.9  | 69.0  | 89.0  | 76.1  | 65.2  | 75.6  |
| Improved HRNet | HRNet-W48       | 640              | 63.8M  | 154.3 | 70.5  | 89.3  | 77.2  | 66.6  | 75.8  |

Figure 5 shows the experimental comparison results of HRNET and Improved HRNET in single-person simple posture, two-person simple posture and multi-person complex posture, which more reflects the superiority of the improved network structure.

5. Conclusions
In this paper, the high resolution network is improved on the basis of high resolution network, and the improved high resolution network structure is introduced in detail. In this paper, in the process of network operation, dual attention mechanism is added to carry out feature extraction without changing the number of channels, focusing on the features with more useful information to reduce the interference brought by irrelevant information. This paper uses Improved HRNet to surpass the previous methods and become the best bottom-up method (AP70.5).
Figure 5 Comparison of experimental results

References
[1] Newell Alejandro, Yang Kaiyu, Deng Jia. Stacked Hourglass Networks for Human Pose Estimation[C]// European Conference on Computer Vision. Springer International Publishing, 2016.
[2] Wei Shih-En, Ramakrishna Varun, Kanade Takeo, et al. Convolutional Pose Machines[J]. IEEE, 2016.
[3] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. arXiv preprint arXiv:1902.09212, 2019.
[4] Jie H, Li S, Gang S, et al. Squeeze-and-Excitation Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, PP(99). Newell Alejandro, Yang Kaiyu, Deng Jia. Stacked Hourglass Networks for Human Pose Estimation[C]// European Conference on Computer Vision. Springer International Publishing, 2016.
[5] Zagoruyko S, Komodakis N. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. 2016.