Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105 (Online)
Volume **7**, Number **4**, November **2022**

High-resolution Human Pose Estimation Method Based on Efficient Convolution

Hai-xia Du

Electronic Engineering College
Heilongjiang University
No.74 Xuefu Road, Harbin, China
hxd1410501@163.com

Hong-Bin Ma*

Electronic Engineering College
Heilongjiang University
No.74 Xuefu Road, Harbin, China
mahongbin@hlju.edu.cn

Zheng Fan

Electronic Engineering College
Heilongjiang University
No.74 Xuefu Road, Harbin, China
fanzheng07@outlook.com

*Corresponding author: Hong-Bin Ma

Received July 4, 2022, revised August 10, 2022, accepted September 25, 2022.

Abstract. *The HigherHRNet network achieves remarkable accuracy in human pose estimation. The network adopts parallel multi-branch complex structure to maintain the network high-resolution state, while a large number of ordinary convolutional multiplication operations generate a large number of redundant feature maps, which both impose a great burden on the number of parameters and computational complexity of the network and affect its deployment in mobile. To address this problem, the network maintains a parallel multi-branch structure while improving the multiplication operation of ordinary convolution, therefore, we introduce the efficient convolutional module, Ghost block. The block uses an appropriate number of convolution filters to process the internal feature mappings, and then applies a series of transformations consisting of linear operations to generate more redundant feature mappings. We design a new bottleneck block, Ghsotbottle, which integrates the SE attention mechanism that obtains more useful channel information by the way of channel assigning different weights. To replace the convolution module and bottleneck module of HigherHRNet network with this pair, therefore, design Lite-HigherHRNet network based on effective convolution for high-resolution human pose estimation. The experiments are validated on the MS COCO2017 dataset, and the experimental results show that the number of network parameters and the computational complexity are effectively reduced while maintaining the detection accuracy.*
Keywords: Human pose estimation, Efficient convolutional module, Channel attention mechanism, Lite-HigherHRNet.

1. **Introduction.** With the rapid development of deep neural networks, human pose estimation technology has also made significant progress and is widely used in human-computer interaction, gesture recognition, pedestrian rediscrimination, target tracking [1, 2] and other fields [3]. For human pose estimation, the state-of-the-art network structures usually include encoders and decoders. Most of the models usually use deep convolutional neural networks

(CNNs) as feature encoders, due to their efficient performance. As far as the decoder is concerned, a heatmap-based [4] approach is usually used to represent the estimated human key points. Since the performance of the heatmap is greatly degraded by the input of low-resolution maps, which greatly affects the performance of key point estimation, it is common to stack up sampling layers to recover the high resolution to improve its localization accuracy. HourGlass [5] improves performance by skip connections to leverage information from feature maps at different scales and has been shown to be effective for human pose estimation. SimpleBaseline [6] adopts a few transposed convolution layers for generating high-resolution representations. HRNet [7] recovers the high resolution from the parallel low resolution, and the whole process always maintains a high-resolution state. ResNet [8] ,VGGNet [9] and other serial networks have excellent performance in image classification, but it is difficult to improve the human posture estimation by widening and deepening the traditional serial networks.

In human posture estimation, the parallel high-resolution network can achieve good performance. In particular, HigherHRNet [10] uses HRNet [7] as the backbone network for feature extraction, and has achieved the best performance in the CrowdPose dataset of more than 150GMACs. However, when the network achieves high accuracy, we also need to consider the network calculation amount and complexity. HRNet parallel multi branch high-resolution network, the low resolution branch captures the context information and the high-resolution branch retains the spatial information. The multi-resolution features are fused through the transmission unit to form a more informative feature map. The parallel high-resolution network structure is also very complex. It requires a lot of computation to recover high resolution from low resolution by up sampling. And a large number of redundant feature maps are generated in the traditional convolution multiplication operation. It affects its deployment in low-performance mobile terminals. We can not ignore its computing efficiency and storage efficiency in order to pursue high performance. Therefore, it is of great significance for us to deeply study an efficient human posture estimation network.

At present, one of the main ways to design efficient networks is to reduce the redundancy by improving the matrix multiplication method of convolution process. For example, MobileNet [11] and ShuffleNet [12] of the classification network. In particular, Huawei proposes that GhostNet [13] has obvious effect in reducing the amount of network parameters. Inspired by this, this paper designed lite-HigherHRNet, which is a high-resolution human pose estimation network based on effective convolution. First, the parallel multi branch network structure for recovering high resolution from low resolution is still maintained. By improving the general convolution multiplication operation method and designing an effective convolution moduleghost module. Ghost module is operated in two steps. Convolution operation can obtain the intrinsic feature information of the input feature layer, and the linear operation layer generates more feature maps from the intrinsic feature map through linear transformation. In particular, the SE [14] attention module is introduced to obtain more effective channel information by weighting the channels to improve the accuracy of the network. Finally, the ghost bottleneck module is proposed, which integrates the effective convolution module, SE attention mechanism and branch structure. Therefore, the ghost bottleneck module is used to replace the HigherHRNet bottleneck module and the ghost module is used to replace its basic module.

2. **Related work.**

2.1. **2D human posture estimation.**

Existing human posture estimation networks are divided into two main categories: top down [6, 15, 16] and bottom-up [17–19]. The top-down approach is to first detect the bounding box from the image using a human detector, and simplify the multiple bounding boxes in the image into a single human pose estimation. For example, Mask R-CNN [20] human pose detection algorithm directly after Faster R-CNN [21] branch, The RoIAlign method is used to make it no longer rounded when downsampling. The error generated by the quantization operation is eliminated and the pixels can be well aligned after downsampling. The G-RMI [22] algorithm follows the common top-down human pose

estimation algorithm, using two independent human detection algorithms and pose estimation algorithms to first detect the human body in the image and then estimate each key point respectively, and finally estimate the overall human pose. The bottom-up approach directly inferred the association information of the joint points and all characters in the image, and grouped the key points according to the association information of the key points of the characters, from which the multi-person pose estimation was performed. OpenPose [23] is a multilevel network that generates a two-branch structure to implement the prediction heat map and the grouping task. OpenPose introduces the part affinity field to compute the 2D vector between two key points to implement the grouping task. PersonLab [24] enables Dilated Resnet networks, which directly learn each pair of key points and their relative 2D displacements to group key points. Wu. et al. [25] proposed that combine vibe with a simple multi-scale abnormal behavior detection algorithm to realize the recognition of human moving targets. A better bottom-up pose estimation method, HRNet [7] network maintains a high resolution multi-branch parallel structured network that allows multi-resolution fusion, which effectively solves the problem of scale variation and achieves good results in bottom-up pose estimation. Higherhrnet [10] uses HRNet as the base network to generate high-resolution feature maps, and by adding deconvolution, higher-resolution feature maps are generated to predict heat maps and grouped using the Associative Embedding method. EfficientHRNet [26] combine the recent advances in model scaling with high-resolution feature representations and creates highly accurate models. LitePose [27] adopt fusion deconv head and large kernel conv to desgin an efficient single-branch architecture for pose estimation.

2.2. **Lightweight network method.** Reduce the number of parameters of the network by compressing the algorithm model, such as Network Pruning [28], Knowledge extraction [29], Quantization [30] etc. But the dominant is to improve the traditional convolutional multiplication operation, the traditional convolutional multiplication operation contains a large number of parameters and floating point operations, convolutional operations include convolutional spatial information calculation and channel information calculation,such as MobileNet [11] and ShuffleNet [12]. these networks reduce the redundancy in matrix vector multiplication, depth-separable convolutional operations can effectively reduce the number of parameters of the network. for example,convolution and skip connection in Squeezenet network can achieve similar results as AlexNet network in image classification and have less number of parameters. MobileNet [11] network uses deep separable convolution, thus sepa rating the fusion of spatial information and channel information,and dividing the convolution process into two steps: channel-by-channel convolution (fusion of spatial information) and point-by-point convolution (fusion of channel information), which greatly reduces the computational effort. ShuffleNet [12] shows that using a large number of convolutions in the network consumes a lot of resources and causes the channels to be full of constraints, which reduces the accuracy of the model to some extent.In order to further reduce the computation, ShuffleNet proposes a channel mixing wash to complete the fusion of information between channels. Huawei proposes the GhostNet [13] network, which designs a new basic neural network unit Ghost block, divides the ordinary convolution in the neural network into two parts, the first part uses an appropriate number of convolution filters to process the internal feature mappings, and then applies a series of transformations consisting of linear operations to generate more redundant feature mappings, which greatly reduces the number of required parameters compared to ordinary convolution.

2.3. **Attention mechanism.** The attention mechanism is widely used in deep learning, such as video saliency display [31], face recognition [32] and other fields, to obtain salient regions by selectively focusing on useful information and ignoring other redundant information. For example, Wu. et al. [31] proposed the use of wavelet transform and feature comparison to achieve the role of attention mechanism to obtain the video salient regions. In the field of image processing, the attention mechanism divides attention into three types: channel domain attention, spatial domain attention, and channel-space hybrid domain attention based on a

mixture of both, and the attention mechanism that is generally not specifically stated is spatial domain attention. And the mechanism of channel-domain attention is to learn more useful features by learning the weights of channels during network transmission.

3. **HigherHRNet network.** The bottom-up HigherHRNet [10] uses HRNet [7] as a feature extraction network to generate high-resolution feature maps, and generates feature maps with resolution by increasing the deconvolution to compensate for the difficulty of detecting human joints with small resolution in images. The high-resolution parallel multi-branch network HRNet consists of four parallel sub-networks, where the feature map resolution of the same subnetwork is kept constant, and the resolution of the feature map generated by downsampling of the parallel sub-network is reduced by half and the number of channels of the network is doubled. The subnets contain in turn the number of modules 4, 3, 2, 1. Each module consists of four residual bottleneck modules and a multi-resolution fusion module. The initial backbone branch uses two convolutions of step 2 to reduce the resolution of the input image to 1/4, the first stage consists of four residual units,each residual unit consists of a bottleneck module with 64 channels,and then the number of channels of the feature map is reduced to C using convolution, and the number of channels of the four resolution convolutions of the parallel branch are C ,2C ,4C ,8C. Generally, a non-normalized Gaussian kernel is used to predict a Gaussian smoothed heatmap, which represents the predicted key point heatmap, In HigherHRnet, the accuracy of the predicted key point heat map is increased by predicting a higher resolution heatmap. Usually, the resolution of the predicted heat map is 1/4 of the input image, and the predicted heatmap generated by HRnet is used as the input, and after deconvolution, a new feature map with two times of the input heat map is generated, so that the predicted heat map of human joints with double-resolution rate can be better predicted to the pose of small and medium people.

4. **Ghost block models.** There is a great similarity between the feature maps output by the convolutional layer, so a large amount of redundancy is generated. The Ghost [13] block is mainly composed of two parts,the convolutional layer and the linear operation layer, which can obtain the intrinsic feature information of the input feature layer, and the linear operation layer can generate more feature maps from the intrinsic feature map by a simple linear transformation. The internal feature maps generated by the linear mapping and the feature maps generated by the convolutional mapping are integrated into the Ghost block, which greatly reduces the number of parameters and the computational effort.

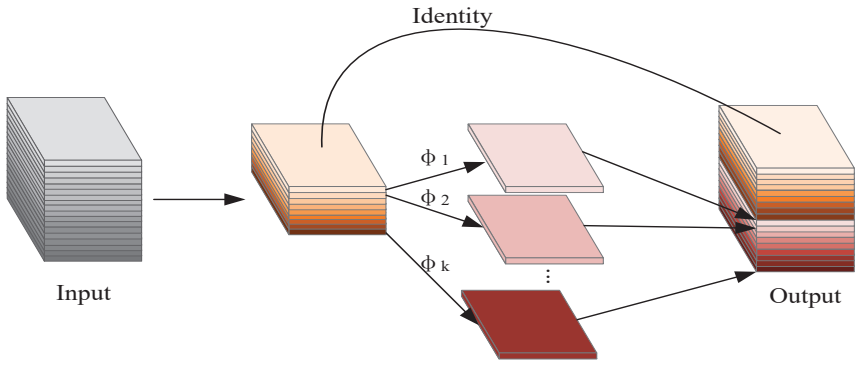

FIGURE 1. Ghost block models

As shown in Figure 1. Ghost module convolution process, the Ghost module first obtains the intrinsic feature map by convolution operation, assuming that the input feature map is c, the number of channels is c,the height of the feature map is h, the width is w, after the convolution f operation, $f \in R^{c\times k\times k\times m}$,the number of channels is m, the convolution kernel is

$k \times k$, The output feature map is $Y_1 \in R^{h' \times w' \times m}$, n denoted as the number of output feature maps, ignoring the bias term, and the convolution operation is expressed as in Equation 1:

$$Y_1 = X * f \tag{1}$$

Second, we obtain an N-dimensional intrinsic feature map by performing a linear operation on the intrinsic feature map to produce more similar feature maps. The input feature map $Y_1 \in R^{h' \times w' \times m}$, output feature map $Y_2 \in R^{h' \times w' \times n}$, and linear operation are expressed in Equation 2:

$$Y_{2ij} = \phi_{ij}\left(Y_{ij}\right), \quad \forall\, i = 1, \ldots, m;\ j = 1, \ldots, s, \tag{2}$$

where: ϕ_{ij} denotes the linear transformation function for the i_{th} linear transformation of the j_{th} feature map. Y_{1i} denotes the hth feature map in the input i_{th} intrinsic feature map. It is worth noting that the computational cost of the linear operation on each channel is much lower than that of the normal convolution operation.

Finally, we stitch the intrinsic feature map of the channel dimension with the feature map generated by the linear operation. As expressed in equation 3:

$$Y = Cat\left(Y_1, Y_2\right) \tag{3}$$

The output feature map is obtained as $Y \in R^{h' \times w' \times n}$, the number of channels is n, and this method can significantly reduce the network operation parameters with less loss of network accuracy.

Assuming that the convolutional kernel size of the network is $k \times k$, the compression factor is s, C_{in} denoted as the number of input channels and C_{out} denoted as the number of output channels, so the ratio of the number of network parameters is expressed in Equation 4:

$$r_p = \frac{\frac{C_{out}}{s} \cdot C_{in} \cdot k \cdot k + (s-1) \cdot \frac{C_{out}}{s} \cdot k \cdot k}{C_{out} \cdot C_{in} \cdot k \cdot k} = \frac{s + C_{in} - 1}{s \cdot C_{in}} \approx \frac{1}{s} \tag{4}$$

The complexity of the network operation is expressed in Equation 5:

$$r_c = \frac{\frac{C_{out}}{s} \cdot H \cdot W \cdot C_{in} \cdot k \cdot k + (s-1) \cdot \frac{C_{out}}{s} \cdot H \cdot W \cdot k \cdot k}{H \cdot W \cdot C_{out} \cdot C_{in} \cdot k \cdot k} = \frac{s + C_{in} - 1}{s \cdot C_{in}} \approx \frac{1}{s} \tag{5}$$

5. **SE attention mechanism.** The filtering function of important channels is achieved by assigning attention weights to the feature map channels to obtain channel information more effectively. The schematic diagram of SE [14] attention module is shown in Figure 2:

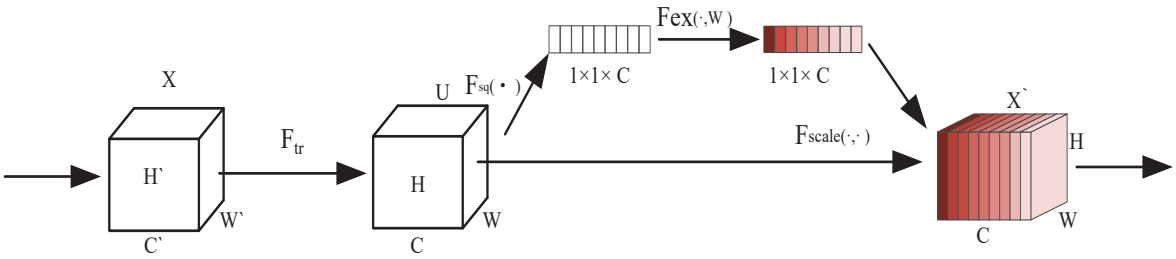

FIGURE 2. Schematic diagram of SE attention module

The input feature map X, W, H, C represents the width, height, and number of channels of the feature map. Firstly, the global average pooling operation is performed on the input F_{TR} channel features to extract the global information of each channel to obtain the $F_{sq}(\cdot)$ in the graph; then the output is connected to the channels of the feature map by two levels of full connection operation to obtain the $F_{ex}(\cdot)$ in the graph.Then the sigmoid normalization operation is performed to normalize to the $0 \sim 1$ range to obtain the final attention weight value of each channel; finally the obtained weight value is multiplied with the initial channel, so that the

network can obtain more effective information about each feature map information. Therefore, by adding the SE attention mechanism to the network, the weights of high information flux channels can be enhanced more effectively and the weights of bottom information channels can be suppressed, thus enhancing the effective utilization of channel information.

Therefore, inspired by the design idea of residual bottleneck structure of Resnet network, we introduce the new Ghsotbottle module, in order to enable the network to obtain more effective information of human keypoints, incorporate the SE attention mechanism, suppress the low utility channel information, weight the high utility channels to obtain more effective utilization. This allows the network to focus on the feature maps that highlight the key points of the human body, making the key points more prominent. The Ghsotbottle module in Stride=1 consists of two Ghost modules stacked: the first Ghost module focuses on the expansion of the channels, and the second Ghost module focuses on the reduction of the number of channels. Similar to the execution residual structure, it is necessary to add the number of channels initially input to the Ghostbottle module and the number of channels passing through the two ghost modules. Therefore, it is necessary to keep the number of channels after passing through the two Ghost modules the same as the input. As for the case where Stride=2, the shortcut path is implemented by a downsampling layer and a depth-wise convolution with Stride=2 is inserted between the two Ghost modules. The specific algorithm is shown in Algorithm 1.

Algorithm 1 Ghostbottle module with SE module structure

Input: $X \in R^{c_{in} \times h \times w}$: represents the output from the previous convolutional layer,the input of this convolutional layer Ghost: Ghsot effective convolution block;SE: channel attention mechanism module;DWconv: deep separable convolution;shortcut: residual structure channel connection;S: representative stride;RES: residual structure input feature map.
Output: $X \in R^{c_{out} \times h \times w}$.

1: $RES \leftarrow X$: The number of channels of the input feature map is C_{in}.
2: $X \leftarrow Ghost1(X)$: The feature map is expanded by Ghost1 convolution block to achieve the number of channels.
3: **if** S>1 **then**:
$X \leftarrow DWconv(X)$: stirde¿1 needs to be downsampled, after DWconv operation.
4: **end if**
5: **if** SE=Ture **then**:
$X \leftarrow SE(X)$: Determines if the SE module is added to weight the channels.
6: **end if**
7: $X \leftarrow Ghost2(X)$: Compress the number of channels of the feature map so that it is the same as the number of input channels.
8: $X \leftarrow X+shortcut(RES)$: The initial feature map of the residual structure is summed with the output feature map, and both of them have the same number of channels.

6. **Lite-HigherHRnet network.** The same HRNet parallel multi-branch network structure is used, and the proposed effective convolution Ghost module is an improvement of the ordinary convolution. Inspired by the construction of bottleneck modules from Resnet network, we also constructed stride=1 and stride=2 Ghostbottle bottleneck modules, The amount of convolution parameters is reduced while balancing the effectiveness of convolution in acquiring information. Therefore, the SE attention mechanism is integrated in the Ghsotbottle bottleneck module to enhance the information exchange between channels by assigning different weights to the channels. In stage1 uses Ghostbottle module with Stride = 2 to replace the bottleneck module stacked by two ordinary convolutions. However, the bottleneck modules of stage2, stage3, and stage4 are basicblock modules replaced by Stride=1 Ghostbottle module. The Conv2D normal convolution of the network in the downsampling transfer module is replaced by the fast convolution Ghost module. By improving the normal convolution of the backbone network, the

computation and the number of parameters of the network are greatly reduced, compared with the original HigherHRNet, the number of parameters of the network is reduced by 42.6% and the computational complexity is reduced by 31.2%, The structure of Lite-HigherHRnet network is shown in Figure 3.

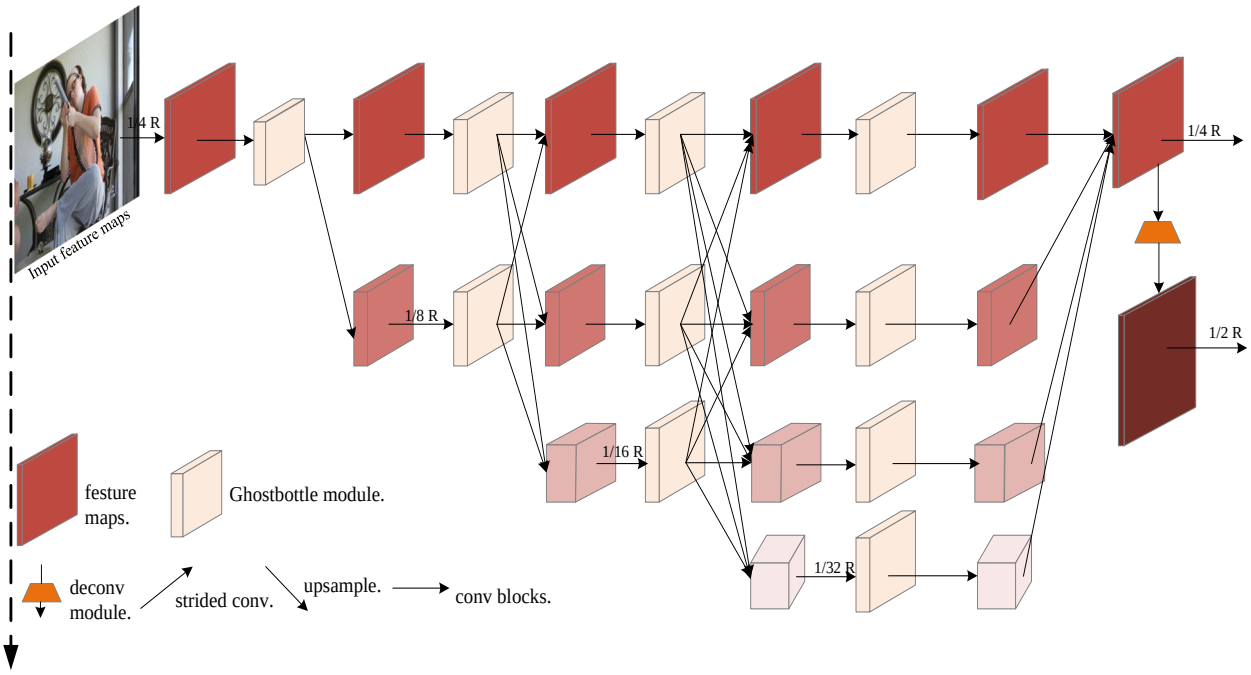

FIGURE 3. Lite-HigherHRnet network architecture.

7. **Experimental results and analysis.**

7.1. **Experimental data and platforms.** Experimental platform: 64-bit Ubuntu 18.04, 24G video memory, two 3090 video cards and Pytorch1.7, python3.8, cuda11.0 experimental platform.

COCO dataset [33]: This experiment is conducted on the Microsoft COCO2017 dataset, which contains 200,000 images and 250,000 pedestrian instances with 17 key point annotations, and the COCO2017 dataset is divided into train/val/test-dev datasets, which have 57k training datasets, 5k evaluation datasets and 20k test datasets, respectively , 5k evaluation datasets and 20k test datasets, respectively. Our experiments are trained on the COCO2017 training level and the obtained results are evaluated on the COCO2017 val evaluation level, and the obtained results are compared with other methods that have excellent results.

7.2. **Evaluation indicators and training strategies.** Objective key point similarity (OKS) is a standard evaluation metric for the COCO dataset, and the OKS is shown in equation 6:

$$OKS = \frac{\sum_i exp\left(-d_i^2/2s^2k_i^2\right)\sigma\left(v_i > 0\right)}{\sum_i \sigma\left(v_i > 0\right)} \tag{6}$$

The following scoring criteria were derived from the AP under the OKS standard evaluation metrics of the COCO dataset.

mAP: the mean value of all APs obtained with 0.05 interval OKS in the range of $[0.50, 0.95]$ as the threshold value.

AP^{50}: accuracy at OKS=0.5.

AP^{75}: accuracy at OKS=0.75.

AP^M: accuracy of human pixels at $[32^2,96^2]$, i.e., medium scale population accuracy.

AP^L: body pixels $> 96^2$, i.e., large scale population accuracy.

8. **Training process.** The data enhancement in this experiment mainly consists of random rotation ($[-30, 30]$), random scale ($[0.75, 1.25]$), and random translation ($[-40, 40]$) cropping operations on the input image of size 512, as well as random flip of the input image. Overscaled heat maps of 128×128 and 256×256 are generated at HigherHRnet by adding deconvolution. Wep use the Adam optimizer and train 300 epochs with a base learning rate of $le-3$, a learning rate of $1e-4$ at the 200_{th} epoch and a learning rate of at the 260_{th} epoch.

8.1. **Experimental results and analysis.** The experimental results of the Lite-HigherHRNet network proposed in this paper are validated on the COCO2017 VAL dataset and compared with the validation results of other state-of-the-art networks, and the obtained results are shown in Table 1 below. In this paper, we propose to use the efficient convolution-Ghost module instead of the ordinary convolution module, by dividing the ordinary convolution into two operations, using the ordinary convolution method to generate the intrinsic intrinsic convolution first, and then generating more redundant feature maps by linear operations. In this paper, a scaling factor of S=2 is used to generate redundant feature maps. According to the formula shown, the number of network parameters and operation more complexity of HigherHRNet will be reduced by 50% in the ideal state, but the SE attention mechanism is added in constructing the Ghostbottle module, which will have an impact on the number of parameters and operation complexity of the network, compared with the HigherHRNet-32 network, the number of parameters in this paper is reduced by 42.6% and the computational complexity is reduced by 31.2%. Compared with the accuracy of HigherHRNet-32, the accuracy is reduced by 1.1%, but compared with HRNet-32, the accuracy is increased by 1.2%. It is found that the accuracy of AP^{50},AP^{75},AP^{M},AP^{L} is partially lower than that of HigherHRNet-32. However, compared with HRNet, the accuracy of AP^{75},AP^{M} is 1.2% and 3%. It can be seen that the network still has a high accuracy in estimating small and medium-sized people and the use of double-resolution heatmap can make up for the error of keypoint prediction. At the same time, the amount of calculation parameters and complexity of the network are greatly reduced. The network always keeps acquiring the feature map at high resolution and incorporates the SE attention mechanism to enhance the effective utilization of channels. The improvement of the convolutional multiplication manipulation can effectively realize the network lightweight, but the addition of the attention mechanism will affect the detection accuracy of the network, so that the network can still achieve good detection results with only a small sacrifice of the detection accuracy while realizing the lightweight.

Table 1. Validation results of the COCO VAL dataset

Method	Input Size	Params	GFLOPS	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
Openpose [23]	\	\	\	61.8	84.9	67.5	57.1	68.2
Hourglass [5]	512	277.2M	206.9	56.6	81.8	61.8	49.8	67.0
PersonLab-101 [24]	1401	68.7M	405.5	66.5	88.0	72.6	62.4	72.3
HRNet-32 [7]	512	28.5M	38.9	64.1	86.3	70.4	57.4	73.9
HigherHRNet-32 [10]	512	28.6M	44.6	66.4	87.5	72.8	61.2	74.2
Lite-HigherHRNet	512	16.4M	30.7	65.3	86.2	71.6	60.4	73.5

8.2. **Experimental visualization and analysis.** This paper is validated on the COCO2017 dataset, which contains 5000 images, and 17 key points of the human body are labeled and the labeled key points are detected. 17 key points are: nose, left eye, right eye, left ear, right ear, left shoulder, right shoulder, left elbow, right elbow, left wrist, right wrist, left crotch, right crotch, left knee, right knee, left ankle, right ankle. In this paper, the network outputs a multi-scale degree heat map for predicting each key point in order to be able to detect the key points of small crowds in images, and a single image and a multi-person image are randomly selected from the COCO2017 VAL dataset to verify the detection results.

As shown in Figure 4, Figure 5, the key point response heat maps of five randomly selected human key points under 128 pixels and 256 pixels in single and multi-person states, respectively, are compared and found that the location of the key points can be predicted more accurately under 256 pixels, and the key points that are obscured in the single-person state accordingly. In the multi-person state, there is a situation where multiple people are obscured from each other and the far field of view, it is found that the key point heat map under 128 pixels will show overlap between heat map and heatmap, and there is a deviation for the positioning of the heatmap, compared to 256 pixels which can accurately predict the location of the key point, and the heatmap for the far field of view condition response is more obvious than 128 pixels. The comparison reveals that the use of dual heatmap prediction can well compensate for the errors in detection for multiple people and distant field of view populations. It indicates that the detection at higher pixels has the ability to resist occlusion and interference.

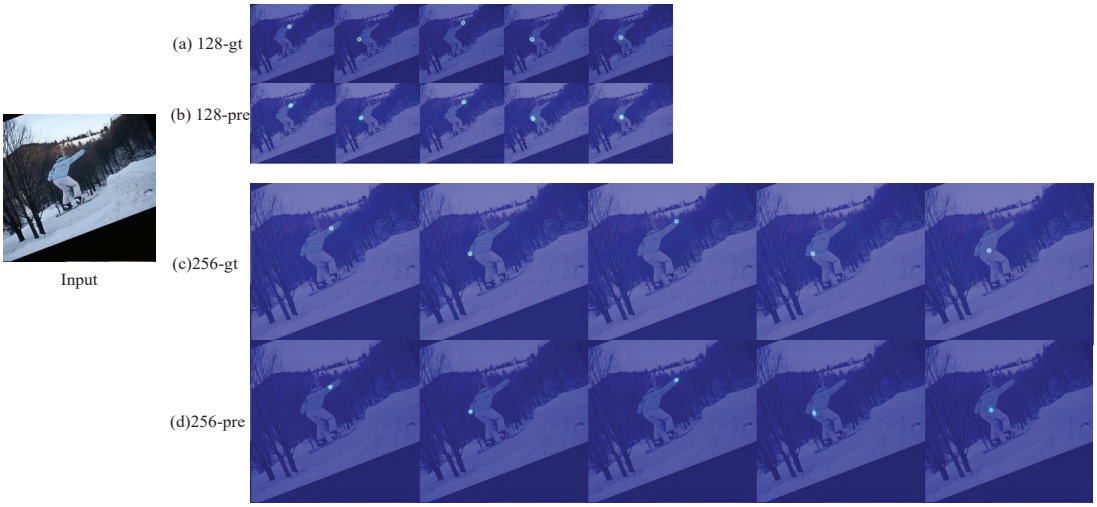

FIGURE 4. Single person key point response characteristics diagram

Note: (a) is the 128-pixel single person labeled keypoint response feature map; (b) is the 128-pixel single person predicted keypoint response feature map; (c) is the 256-pixel single person labeled keypoint response feature map; (d) is the 256-pixel single person predicted keypoint response feature map.

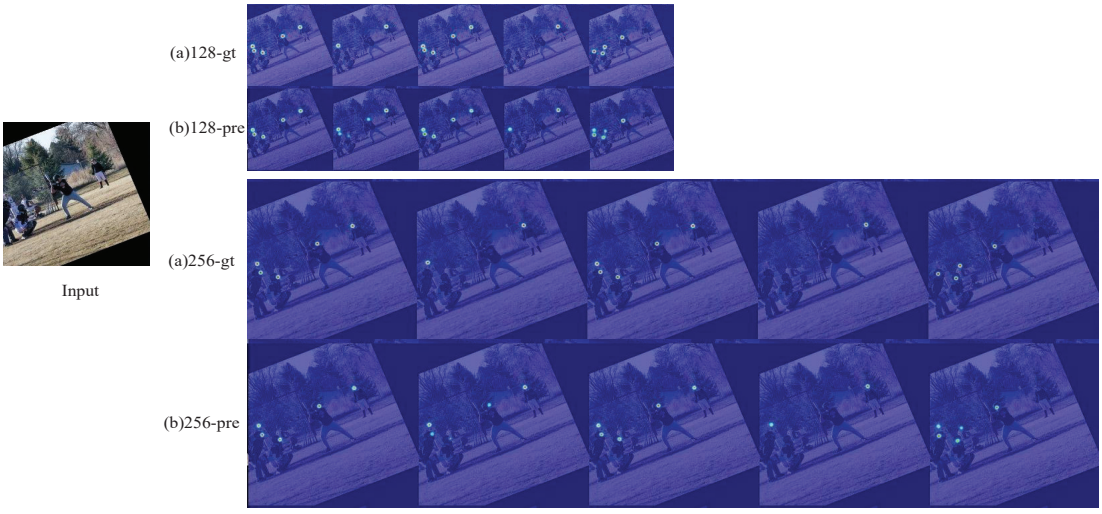

FIGURE 5. Multi-person critical point response characteristics diagram

Note: (a) is the 128-pixel multiplayer labeled keypoint response feature map; (b) is the 128-pixel multiplayer predicted keypoint response feature map; (c) is the 256-pixel multiplayer labeled keypoint response feature map; (d) is the 256-pixel multiplayer predicted keypoint response feature map.

In this paper, the human pose estimation graph can be obtained by connecting the key points according to the detection through the visualization tool, as shown in figure 6 COCO dataset human pose estimation can be viewed.

FIGURE 6. Human posture estimation results for the COCO dataset

(a) indicates the result of human pose estimation in single person state; (b) (c) two figures appear multi-person condition, it is obvious to find that in (b) figure the persons are obscured from each other, which affects the detection of key points of the human body as a whole, and can only estimate the pose of the detected part. In figure (c), there is no serious occlusion between multiple people, and the pose of the human body in the far field of view can also be estimated.(d) shows a self-occlusion situation where the key points of the human part are not detected, which affects the overall pose estimation, and the pose estimation is not possible due to the incomplete and blurred state of the human in the far field of view. In this paper, we use Ghost module to achieve a lightweight network but at the same time reduce the detection accuracy of the network, but the network maintains a high-resolution extraction of feature maps and output dual-resolution prediction heat map, when there are multiple people blocking each other, the small field of view of the human body can also detect the key points of the human body to estimate the human body.

9. **Conclusion.** HigherHRNet [10] uses a high-resolution parallel network as the backbone network for feature extraction and achieves excellent performance in human pose estimation, but the large number of parameters of the network limits its deployment at the network side. To address this issue, we still use the network structure of high-resolution branching to maintain the network feature extraction accuracy, because the normal convolution process generates a large number of redundant maps, and therefore, the multiplication operation of convolution is improved to reduce the number of parameters generated by the convolution process. Designing the effective convolution Ghost module. It implements the convolution process in two steps, first using ordinary convolution to generate the intrinsic feature map, and then a linear operation to generate the redundant feature map to reduce the number of outputs of the convolution process. The basic bottleneck module of HigherHRNet is replaced by the Ghostbottle

module, but the Ghostbottle module is built based on two effective convolutional Ghost modules. While achieving a lightweight network, the Ghostbottle module incorporates a channel attention mechanism to enhance its accuracy by effectively utilizing important channel information. The results from the experiments show that, compared to the HigherHRNet network. the amount of network parameters is reduced by 42.6% and the computational complexity is reduced by 31.2%. Although it has an impact on the detection accuracy of the network, the network can still maintain a good detection accuracy by using the output dual-resolution prediction heat map, which can achieve accurate detection of key points of small and medium-sized populations in the network, thus proving the research to be of great significance. We can build simpler network structures with effective convolutional multiplication operations. Combining the two ways of improving the network structure and improving the convolution operation, and studying them in depth, we can achieve the light weight of human pose estimation and solve its deployment problem.

REFERENCES

[1] J. Gao, H. Zou, F. Zhang, and T. Wu, "An intelligent stage light-based actor identification and positioning system," *International Journal of Information and Computer Security*, vol. 18, no. 1/2, pp. 204–218, 2022.

[2] F. Zhang, T. Y. Wu, J. S. Pan, G. Ding, and Z. Li, "Human motion recognition based on svm in vr art media interaction environment," *Human-centric Computing and Information Sciences*, vol. 9, no. 1, p. 40, 2019.

[3] J. Yu, Y. Rui, and D. Tao, "Click prediction for web image reranking using multimodal sparse coding," *IEEE Transactions on Image Processing*, vol. 23, no. 5, pp. 2019–2032, 2014.

[4] Y. Li, S. Yang, P. Liu, S. Zhang, Y. Wang, Z. Wang, W. Yang, and S.-T. Xia, "SimCC: a Simple Coordinate Classification Perspective for Human Pose Estimation," *arXiv e-prints*, p. arXiv:2107.03332, Jul. 2021.

[5] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision*. Springer, pp. 483–499, 2016.

[6] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 466–481, 2018.

[7] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5693–5703, 2019.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[9] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv e-prints*, p. arXiv:1409.1556, Sep. 2014.

[10] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang, "Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5386–5395, 2020.

[11] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, 2018.

[12] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 116–131, 2018.

[13] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "Ghostnet: More features from cheap operations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1580–1589, 2020.

[14] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, 2018.

[15] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7103–7112, 2018.

[16] C. Yu, B. Xiao, C. Gao, L. Yuan, L. Zhang, N. Sang, and J. Wang, "Lite-hrnet: A lightweight high-resolution network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10 440–10 450, 2021.

[17] S. Jin, W. Liu, E. Xie, W. Wang, C. Qian, W. Ouyang, and P. Luo, "Differentiable hierarchical graph grouping for multi-person pose estimation," in *European Conference on Computer Vision*. Springer, pp. 718–734, 2020.

[18] S. Kreiss, L. Bertoni, and A. Alahi, "Pifpaf: Composite fields for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11 977–11 986, 2019.

[19] Z. Geng, K. Sun, B. Xiao, Z. Zhang, and J. Wang, "Bottom-up human pose estimation via disentangled keypoint regression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14 676–14 686, 2021.

[20] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961–2969, 2017.

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Proceedings of the 28th International Conference on Neural Information Processing Systems(NIPS 2015)*, pp. 91–99, 2015.

[22] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4903–4911, 2017.

[23] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7291–7299, 2017.

[24] G. Papandreou, T. Zhu, L.-C. Chen, S. Gidaris, J. Tompson, and K. Murphy, "Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 269–286, 2018.

[25] G. Ke, R.-S. Chen, Y.-C. Chen, Y.-X. Hu, and T.-Y. Wu, "Simple multi-scale human abnormal behaviour detection based on video," *International Journal of Information and Computer Security*, vol. 17, no. 3-4, pp. 310–320, 2022.

[26] C. Neff, A. Sheth, S. Furgurson, J. Middleton, and H. Tabkhi, "Efficienthrnet: Efficient and scalable high-resolution networks for real-time multi-person 2D human pose estimation," *Journal of Real-Time Image Processing*, vol. 18, no. 4, pp. 1037–1049, 2021.

[27] Y. Wang, M. Li, H. Cai, W.-M. Chen, and S. Han, "Lite pose: Efficient architecture design for 2d human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13 126–13 136, 2022.

[28] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf, "Pruning Filters for Efficient ConvNets," *arXiv e-prints*, p. arXiv:1608.08710, Aug. 2016.

[29] T. Garbay, O. Chuquimia, A. Pinna, H. Sahbi, X. Dray, and B. Granado, "Distilling the knowledge in cnn for wce screening tool," in *2019 Conference on Design and Architectures for Signal and Image Processing (DASIP)*. IEEE, pp. 19–22, 2019.

[30] K. Wang, Z. Liu, Y. Lin, J. Lin, and S. Han, "Haq: Hardware-aware automated quantization with mixed precision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8612–8620, 2019.

[31] F. Zhang, T.-Y. Wu, and G. Zheng, "Video salient region detection model based on wavelet transform and feature comparison," *EURASIP Journal on Image and Video Processing*, vol. 2019, no. 1, pp. 1–10, 2019.

[32] T. M. Liang, M. S. Chiu, Y. C. Wu, M. T. Yeh, C. H. Hsu, and Y. N. Chung,"Applying Image Processing Technology to Face Recognition ," *Journal of Information Hiding and Multimedia Signal Processing*, vol. 13, no. 2, pp. 106-112, 2022.

[33] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*. Springer, pp. 740–755, 2014.