Lightweight and Efficient High-Resolution Network for Human Pose Estimation

Jiarui Liu[1], Xiugang Gong[2]*, Qun Guo[3]
School of Computer Science and Technology, Shandong University of Technology, Zibo, Shandong, 255000, China

***Abstract*—To address the challenges of high parameter quantities and elevated computational demands in high-resolution network, which limit their application on devices with constrained computational resources, we propose a lightweight and efficient high-resolution network, LE-HRNet. Firstly, we designs a lightweight module, LEblock, to extract feature information. LEblock leverages the Ghost module to substantially decrease the number of model parameters. Based on this, to effectively recognize human keypoints, we designed a Multi-Scale Coordinate Attention Mechanism (MCAM). MCAM enhances the model's perception of details and contextual information by integrating multi-scale features and coordinate information, improving the detection capability for human keypoints. Additionally, we designs a Cross-Resolution Multi-Scale Feature Fusion Module (CMFFM). By optimizing the upsampling and downsampling processes, CMFFM further reduces the number of model parameters while enhancing the extraction of cross-branch channel features and spatial features to ensure the model's performance. The proposed model's experimental results demonstrate accuracies of 69.3% on the COCO dataset and 88.7% on the MPII dataset, with a parameter count of only 5.4M, substantially decreasing the number of model parameters while preserving its performance.**

***Keywords*—Human pose estimation; model lightweighting; Ghost module; attention mechanism; multi-scale feature fusion**

I. INTRODUCTION

Human pose estimation, as a core topic in the field of computer vision, aims to recognize and locate keypoints of the human body from images or videos. The key to this task lies in accurately understanding and analyzing human posture and movement, which is crucial for computer vision to comprehend and process complex scenes. Human pose estimation plays an important role in numerous application areas, such as sports analysis, human-computer interaction, and security monitoring [1] [2].

The research on human pose estimation has evolved from early model-based and traditional learning algorithm-based methods, such as graphical models and handcrafted feature extraction [3], to recent deep learning-based methods. Deep learning methods, particularly Convolutional Neural Network (CNN) [4], have significantly improved the accuracy and robustness of techniques for recognizing and locating key-points of the body. This heatmap-based approach effectively handles complex scenes and multi-person pose estimation tasks. Since heatmaps can intuitively represent the positional probability of each keypoint, the model can accurately recognize key points even in cases of partial occlusion or overlap of the human body.

In recent years, numerous classic human pose estimation algorithms have emerged [5][10], achieving significant advancements in recognizing and locating human keypoints in images or videos. Particularly, High-resolution network (HRNet) [11], with their unique network structure and high-resolution feature representation capabilities, can achieve effective human pose estimation while maintaining high accuracy, making them widely applicable in various scenarios. However, due to their complex network structure and large number of parameters and high computational demands, high-resolution networks face difficulties when deployed on resource-constrained devices. Lite-HRNet [12] effectively reduces the model's parameter count by incorporating a Conditional Channel Weighting module. Dite-HRNet [13] introduces dynamic lightweight processing, multi-scale context information extraction, and long-range spatial dependency modeling in high-resolution networks, ensuring model performance with lower parameters. X-HRNet [14] incorporates Spatially Unidimensional Self-Attention (SUSA) for lightweight processing, significantly reducing model parameters without compromising accuracy. These methods have made significant progress in model lightweighting. However, human pose estimation is a task highly sensitive to positional information, and lightweighting high-resolution networks can lead to the loss of critical human keypoint positional information. During multi-scale feature fusion, frequent upsampling and downsampling operations introduce a computational burden. Furthermore, downsampling reduces the spatial detail in feature maps, which is difficult to recover during upsampling.

In response to the issues mentioned above, we propose a lightweight and efficient high-resolution network, LE-HRNet. We utilize the Ghost module to reduce the model's parameter count and introduce a novel attention mechanism in LE-HRNet to enhance the detection of keypoint positional information. This approach ensures model performance while lowering both the parameter count and computational load. Additionally, we optimize the multi-scale feature fusion stage to further decrease computational demands and enhance the extraction of channel and spatial dimensional feature information. The main contributions of this paper are summarized as follows:

- We designed a lightweight module, LEblock, for extracting feature information. We used the Ghost module instead of standard convolution to reduce the model's parameter count, and designed a Multi-Scale Coordinate Attention Mechanism to enhance the

*Corresponding Author.

detection capability of human key points, ensuring the model's performance.

- We optimized the multi-scale feature fusion stage and proposed a Cross-Resolution Multi-Scale Feature Fusion Module. This module optimizes the upsampling and downsampling processes, and by learning cross-branch channel information and spatial features, it ensures the model's performance while further reducing the model's parameter count.
- We conducted experimental validation on the COCO dataset and MPII dataset to demonstrate the effectiveness of the proposed method.

The structure of this paper is organized as follows: Section II introduces the main methods proposed in this paper. Section III conducts experimental verification on the COCO and MPII datasets and analyzes the experimental results. Section E summarize the research results and discuss the future work of LE-HRNet.

II. PROPOSED METHOD

HRNet is widely used for visual tasks that require detailed features, such as human pose estimation and semantic segmentation. Unlike most existing methods that recover high-resolution features from low-resolution features, HRNet connects high-resolution to low-resolution subnets in parallel, maintaining high-resolution feature representation throughout the entire network. It extracts feature information using residual block [15] and achieves multi-scale information exchange through multi-scale feature fusion. This design allows HRNet to effectively retain and utilize high-resolution detailed feature information, enabling more precise capture of image details during the multi-scale feature fusion process. HRNet processes multiple resolution feature maps in parallel within its structure and facilitates information interaction and fusion between feature maps of different resolutions, allowing it to simultaneously acquire local detailed information and high-level semantic information.

Although HRNet has made significant progress in terms of performance, its high parameter count and computational demands make it challenging to apply to devices with constrained computational resources, hindering its practical application value. To address this issue, we propose a lightweight high-resolution network, LE-HRNet, based on HRNet, aimed at human pose estimation. Its structure is shown in Fig. 1.

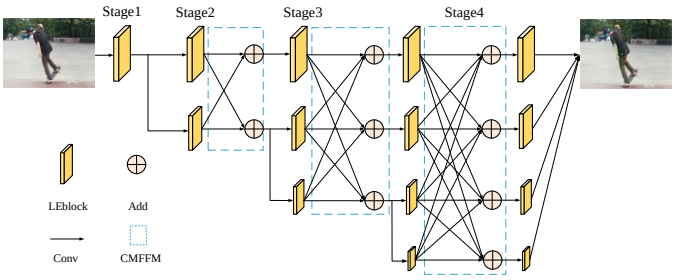


Fig. 1. LE-HRNet structure diagram.

As shown in Fig. 1, we use LEblock instead of residual blocks for feature extraction. This block effectively lowers the number of model parameters and computational demands while minimizing performance degradation, ensuring the model's ability to detect keypoints. Between different resolution feature maps, we optimize the sampling process and use CMFFM for multi-scale feature fusion. This process further reduces the parameter count and computational load, and enhances performance by learning cross-resolution channel and spatial information.

A. ELblock

HRNet uses residual block as the feature extraction module. While residual block effectively enhance the model's feature representation capability, they also bring a large number of parameters and computational load. Therefore, this paper proposes a lightweight block, LEblock, based on the residual block. The structure of LEblock is shown in Fig. 2. We substitute the conventional 3×3 convolution in the residual block with the Ghost module [16] to reduce the number of model parameters and computational overhead. To minimize performance loss and enhance the detection capability of human keypoints while lightweighting the model, we designed and added a Multi-Scale Coordinate Attention Mechanism, which improves the detection of human keypoints with a smaller computational load.

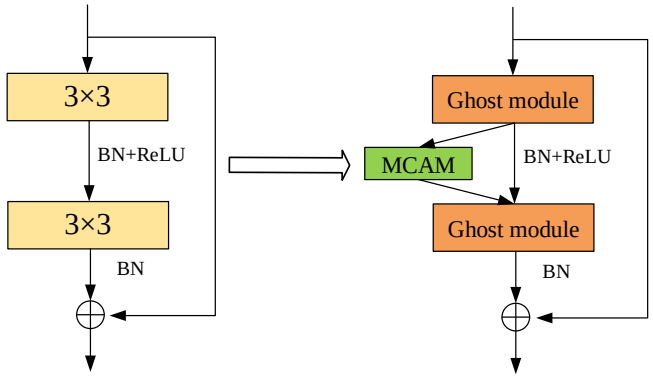


Fig. 2. The left is the residual block and the right is the ELblock.

B. Ghost Module

In traditional convolution operations, a large number of parameters and computations are generated, many of which are redundant. Ghost module optimizes the convolution process to obtain more image features with fewer parameters, thereby achieving model lightweighting. Ghost module decomposes the standard convolution process into three main steps: first, the number of feature map channels is reduced to generate the initial feature map; second, the initial feature map undergoes linear transformations to generate Ghost feature maps; finally, the initial feature map and the generated Ghost feature maps are concatenated to form the final output feature map. The specific transformation process of the Ghost module is shown in Fig. 3.

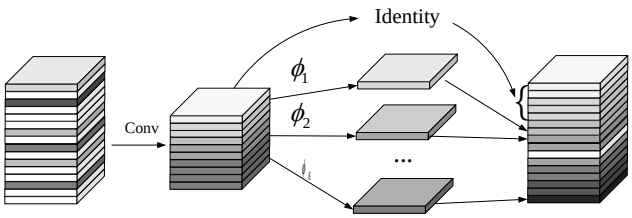


Fig. 3. The Ghost module.

Assuming the input feature map size is $C\times H\times W$, the number of output channels is N, and the convolution kernel size is $k\times k$, the number of parameters for standard convolution is:

$$Param_{conv} = k\times k\times C\times N \tag{1}$$

Among them, C and N are usually quite large, which results in a large number of parameters.

To address this, Ghost Convolution compresses the number of channels in the first step, reducing the channels to $m = N/s$. Next, linear transformations are used to generate Ghost feature maps with the number of channels $m\times(s-1) = N/s$. Finally, the feature maps obtained from the first two steps are concatenated along the channel dimension using an identity operation, resulting in an output feature map with N channels. Assuming $k = d$ and s is much smaller than C, the number of parameters for Ghost module can be calculated as follows:

$$Param_{Ghost} = k\times k\times C\times m + (s-1)\times d\times d\times m \tag{2}$$

Compared to standard convolution, the parameter compression ratio r_{param} for Ghost Convolution is:

$$r_{param} = \frac{Param_{conv}}{Param_{Ghost}} \approx \frac{s\times C}{s+C-1} \approx s \tag{3}$$

Through the Ghost module, the number of parameters can be reduced by a factor of s, achieving a significant reduction in parameter count.

C. Multi-Scale Coordinate Attention Mechanism

Although using the Ghost module to replace 3×3 convolution can decrease the parameter count and enhance computational efficiency, it also weakens the model's feature representation capability, leading to performance degradation and affecting the final prediction results. To enhance the detection capability of the model, attention mechanisms are commonly employed. SE (Squeeze-and-Excitation) [17] and ECA (Efficient Channel Attention) [18] enhance feature representation by re-weighting the channels of feature maps. The SE block integrates information between channels through global average pooling and fully connected layers, while the ECA enhances features through local cross-channel interactions. CBAM (Convolutional Block Attention Module) [19] combines channel attention and spatial attention, extracting global feature information through global average pooling and max pooling, capturing inter-channel dependencies and important spatial information, further enhancing feature representation and model performance. Human pose estimation is a task highly sensitive to positional information, making this information crucial. Coordinate Attention Mechanism[20] encodes spatial information by performing global average pooling in horizontal and vertical directions on the input feature map, and then fuses channel information to generate coordinate attention weights, re-weighting the input feature map. This not only enhances channel feature representation but also captures critical spatial information, thereby improving the model's feature expression capability and overall performance. However, Coordinate Attention Mechanism promotes channel fusion by using channel dimension reduction and expansion, which, although reducing the number of parameters, results in the loss of feature information during the reduction process. Additionally, 1×1 convolution are limited in their ability to extract local feature information and overlook the positional dependencies between different keypoints. Therefore, we propose a Multi-Scale Coordinate Attention Mechanism (MCAM) rove the model's ability to detect human keypoints. The structure of MCAM is shown in Fig. 4.

MCAM enhances the semantic and spatial information of feature maps by using feature grouping, parallel sub-branches, and multi-scale feature learning, producing better pixel-level attention without losing channel dimension information. For the input feature map, to avoid performance loss caused by channel dimension reduction, the input feature map is divided

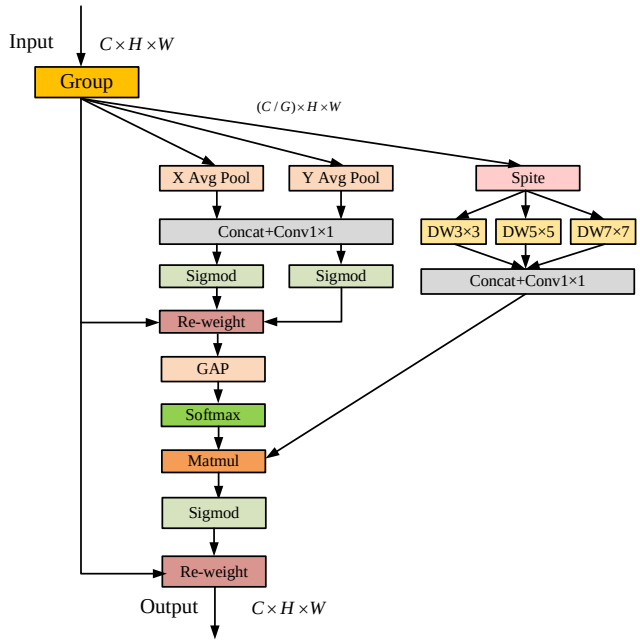


Fig. 4. The structure of multi-scale coordinate attention mechanism.

into multiple sub-feature maps to extract different semantic information. Assuming the input feature map is F_{input} and the output feature map is F_{output}, the generated multiple sub-feature maps are as follows:

$$F_{submap1}, F_{submap2}, ..., F_{submapg} = Group(F_{input}) \tag{4}$$

For each sub-feature map, three parallel sub-branches are used to extract coordinate position information and multi-scale

feature information. The first two sub-branches use operations similar to the coordinate attention mechanism, performing global average pooling in the vertical and horizontal directions to generate direction-aware attention maps. Then, using concatenation and 1×1 convolution, channel fusion is promoted without channel dimension reduction. Finally, the fused coordinate information feature maps are output through the sigmoid activation function. The formula is as follows:

$$\begin{cases} z_x = f_x^{GAP}(F_{submapi}) \\ z_y = f_y^{GAP}(F_{submapi}) \end{cases} \tag{5}$$

$$F_{mid} = f_{conv1\times1}([z_x, z_y]) \tag{6}$$

$$\begin{cases} g_x = \sigma(F_{mid}) \\ g_y = \sigma(F_{mid}) \end{cases} \tag{7}$$

$$F_{Coord\_i} = F_{submapi} \times g_x \times g_y \tag{8}$$

In another sub-branch, we split into three branches and apply depthwise convolution with different kernel sizes of 3×3, 5×5, and 7×7. Smaller kernels can extract local feature information, while larger kernels, due to their larger receptive fields, can more easily extract relevant features between different keypoints. By integrating multi-scale feature information from local to global, we can enhance the model's ability to detect human keypoints in complex scenes, further improving the overall performance and robustness of the model. Finally, multi-scale feature maps are generated through concatenation and 1×1 convolution. The formula is as follows:

$$\begin{cases} F_3, F_5, F_7 = f_{splite}(F_{submapi}) \\ F_{multi} = f_{1\times1}([f_{DW3\times3}(F_3), f_{DW5\times5}(F_5), f_{DW7\times7}(F_7)]) \end{cases} \tag{9}$$

The feature map with coordinate information is modeled in the channel dimension through GAP and Softmax, and fused with the multi-scale feature map to ultimately output the feature map that integrates multi-scale feature information and positional information. The formula is as follows:

$$\begin{cases} \omega_i = \sigma(f_{GAP}(f_{soft\max}(F_{Coord\_i})) \otimes F_{mulit}) \\ F_{outputi} = \omega_i \times F_{submapi} \\ F_{output} = [F_{output1}, F_{output2}, ..., F_{outputg}] \end{cases} \tag{10}$$

Compared to existing attention mechanisms, MCAM not only offers higher computational efficiency but also avoids the compression and expansion in the channel dimension, which reduces the loss of feature information. MCAM effectively enhances the detection capabilities for human keypoints by integrating coordinate positional information with multi-scale feature information.

D. Cross-Resolution Multi-Scale Feature Fusion Module

HRNet enhances the network's understanding of feature maps from local to global by integrating multi-scale feature information through a process of multi-scale feature fusion, improving recognition accuracy and adaptability. However, frequent upsampling and downsampling operations increase the computational burden. Additionally, these sampling processes can lead to the loss of spatial feature information, adversely affecting the model's performance.

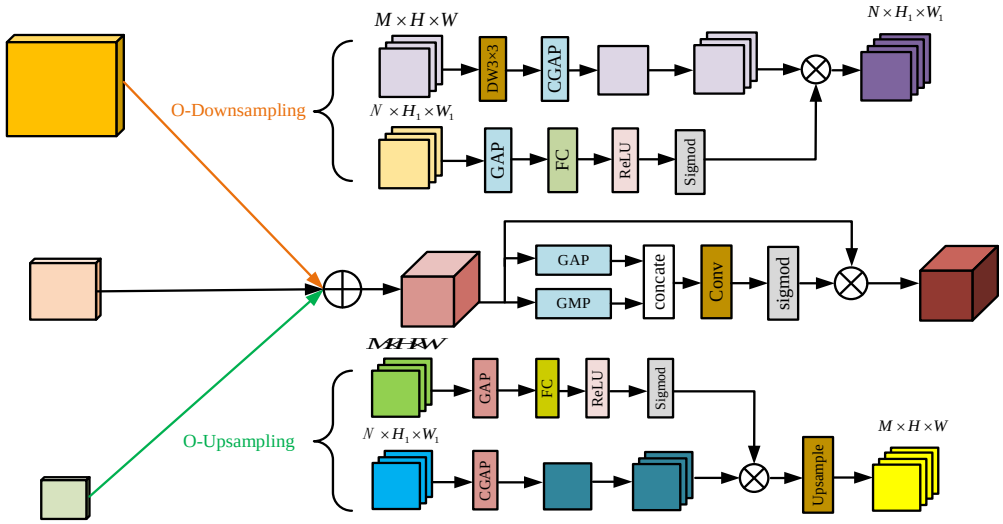


Fig. 5. The structure of cross-resolution multi-scale feature fusion module.

To address this, this paper optimizes the sampling process and proposes a cross-resolution multi-scale feature fusion module, as shown in Fig. 5. Taking the downsampling process as an example, a depthwise convolution with a stride of 2 is applied to the high-resolution branch feature map to reduce its computational load. Global average pooling is then performed along the channel dimension, reducing the number of channels to 1, ensuring that the information of each channel is uniformly preserved in a single feature map, thereby maintaining the essential information of the channels. The pooled feature map is then duplicated *N* times, where *N* is the number of channels of the low-resolution feature map.

For the low-resolution branch feature map, global average pooling is performed along the spatial dimension, and the channel weights are generated through a fully connected layer,

ReLU, and Sigmoid functions. The channel weight information is multiplied with the newly generated low-resolution feature map to achieve fusion, resulting in the optimized downsampled feature map. This process reduces the computational load during downsampling and significantly enhances the response to important feature channels while suppressing the response to unimportant channels by learning cross-resolution channel weight information. The optimized downsampling computation formula is as follows:

$$\begin{cases} \omega = \sigma(\mathrm{Re}\,LU(f_{FC}(f_{GAP}(F_{low-R})))) \\ F_{down} = \omega \otimes f_{copy}(f_{CGAP}(f_{DW3\times3}(F_{high-R}))) \end{cases} \tag{11}$$

Assuming the size of the high-resolution feature map is $M \times H \times W$ and the size of the low-resolution feature map is $N \times H_1 \times W_1$, the compression ratio r_{flops} of the computational load in the downsampling process is:

$$r_{flops} = \frac{Flops_{new}}{Flops_{original}} = \frac{H_1 \times W_1 \times M \times 3 \times 3}{H_1 \times W_1 \times N \times M \times 3 \times 3} = \frac{1}{N} \tag{12}$$

From the formula, it can be seen that this optimization effectively reduces the computational load in the downsampling process. The upsampling process is similar to the downsampling process. The processing flow for upsampling is as follows:

$$\begin{cases} \omega = \sigma(\mathrm{Re}\,LU(f_{FC}(f_{GAP}(F_{high-R})))) \\ F_{up} = f_{upsamp}(\omega \otimes f_{copy}(f_{CGAP}(F_{low-R}))) \end{cases} \tag{13}$$

We optimize the upsampling and downsampling processes in two main steps. The first step involves channel-wise aggregation compression, where information from different channels is merged into one channel. This representation encapsulates key information from multiple channels, resulting in a comprehensive feature representation. The second step focuses on learning cross-resolution channel weight information and using these weights to model the to-be-sampled feature maps along the channel dimension. This optimizes feature selection and reorganization, enhancing the model's representational ability and processing efficiency. These steps not only effectively reduce the computational load but also enhance the model's adaptability and sensitivity to features of different scales. By adjusting channel weights, we can provide varying degrees of emphasis on features at different levels, thus balancing detail and global information better during the upsampling or downsampling processes.

To retain more spatial information and enhance the ability to extract spatial features, we draw on the ideas of CBAM and use a spatial attention mechanism to extract spatial information. First, global average pooling and global max pooling are used to capture the average feature information and salient features of the feature map, respectively. These pooled features are then combined to form a more comprehensive feature representation. A 7×7 convolution is applied to further extract richer spatial feature information, and a Sigmoid activation function is used to generate spatial weights. These spatial weights are fused with the original feature map to produce a weighted and enhanced feature map, effectively preserving and highlighting the spatial details in the feature map.

III. EXPERIMENT

A. Datasets and Evaluation Metric

COCO (Common Objects in Context) [21] is a large-scale dataset widely used in computer vision, particularly suitable for human pose estimation, object detection, and image segmentation. This dataset offers rich scene complexity and extensive category coverage, including over 200,000 images and 250,000 human-annotated object in-stances. For human pose estimation, COCO meticulously annotates 17 keypoints cov-ering the major joints and parts of the body, making it an essential resource for re-searching and developing advanced human pose recognition algorithms.

MPII [22] dataset is a large-scale dataset focused on human pose estimation, containing over 25,000 images spanning 410 types of activities. Each image is detailed and annotated with 16 human body keypoints, including the head, neck, shoulders, elbows, hands, hips, knees, and feet. These images are derived from everyday life scenes, encompassing both individual and multi-person interactions, making MPII not only extensively used in academic research but also crucial for developing practical application algorithms in pose recognition.

In the COCO dataset, the performance of human pose estimation is primarily assessed using Object Keypoint Similarity (OKS). OKS is an evaluation metric that compares the similarity between predicted keypoints and true keypoints. The formula for calculating OKS is as follows:

$$OKS = \frac{\sum_i \exp\left[\frac{-d_i^2}{2s^2k_i^2}\delta(v_i > 0)\right]}{\sum_i \delta(v_i > 0)} \tag{14}$$

Where d_i is the Euclidean distance between the ground truth and predicted keypoint i; k is the constant for keypoint i; s is the scale of the ground truth object; v_i is the is the ground truth visibility flag for keypoint i; $\delta(v_i > 0)$ is the Dirac-delta function which computes as 1 if the keypoint i is labeled, otherwise 0.

OKS can be understood as a normalized measure of the error between the predicted and true annotations for each keypoint, which takes into account the size of the human body and the specific sensitivity of each keypoint. Based on OKS, the COCO dataset also calculates Average Precision (AP) at multiple thresholds, ranging from OKS=0.50 (looser matching) to OKS=0.95 (very strict matching).

MPII uses PCK (Percentage of Correct Keypoints) as the main metric to evaluate model performance. PCK measures the percentage of predicted keypoints that match the true keypoints within a certain distance threshold. The specific calculation formula is as follows:

$$PCK = \frac{1}{n}(\sum_{i=1}^{n} \delta(dist(p_i, q_i) \le \alpha \cdot \max(h, w)) \tag{15}$$

Where n is the number of keypoints; p_i is the predicted position of the i-th keypoint; q_i is the actual position of the i-th keypoint; $dist(p_i, q_i)$ is the distance between the pre- dicted keypoint p_i and the actual keypoint q_i ; α is a predefined threshold, and ; $dist(p_i, q_i) < \alpha$ means the prediction is considered accurate.

B. Experimental Setup

The experimental setup for this paper is as follows: Intel(R) Xeon(R) Silver 4310 CPU @ 2.10GHz, 64GB RAM, two RTX A5000 GPU with 24GB VRAM each, Ubuntu 22.04.3 LTS, Python 3.8. The deep learning framework used is Pytorch 3.9, with CUDA 11.5 for accelerated computing.

When training on the COCO training set, images from the COCO training set are cropped and scaled to a fixed size of 256×192. Adam is used as the optimizer during network training, with an initial learning rate of 0.001. The learning rate is reduced to 0.0001 at the 170th epoch, and then to 0.00001 at the 210th epoch, with a total of 230 epochs of training. During training, random image rotation and horizontal flipping are also used for data augmentation. When training on the MPII dataset, cropped images are uniformly scaled to a fixed size of 256×256. Other training details are the same as those for the COCO dataset, using the same parameter settings and experimental environment.

C. Result and Analysis

The performance comparison of LE-HRNet with other human pose estimation algorithms on the COCO validation set, with an input size of 256×192, is shown in Table I. As seen from the table, compared to HRNet, LE-HRNet reduces the number of parameters and computational load by 81.1% and 78.7% respectively, while the AP only decreases by 4.1%. LE-HRNet achieves a significant reduction in model parameters and computational load with minimal performance loss, maintaining a balance between model performance and parameter size. Compared to Hourglass and CPN, LE-HRNet has lower parameters and higher performance. Compared to SimpleBaseline, LE-HRNet's AP is only 1.1% lower, but its number of parameters is much lower than SimpleBaseline. Compared to lightweight models like MobileNetV2[23] and ShuffleNetV2[24], LE-HRNet's AP is higher by 4.7% and 9.4% respectively, and LE-HRNet also has an advantage in terms of parameters and computational load. Compared to even smaller lightweight models like Lite-HRNet, Dite-HRNet, and X-HRNet, LE-HRNet has more parameters, but its AP is higher by 2.1%, 1.0%, and 1.9% respectively. Unlike these models which aggressively pursue lightweight design, LE-HRNet focuses more on balancing parameter size and performance, ensuring model performance while reducing the number of parameters.

With an input size of 384 × 288, the performance comparison on the COCO test set is shown in Fig. 6. LE-HRNet achieved an AP score of 72.1, outperforming other networks while maintaining a balance between performance and computational load.

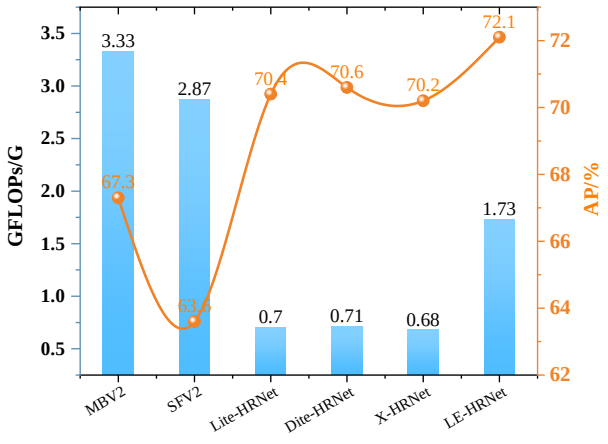


Fig. 6. Performance comparison of different algorithms on the COCO test set.

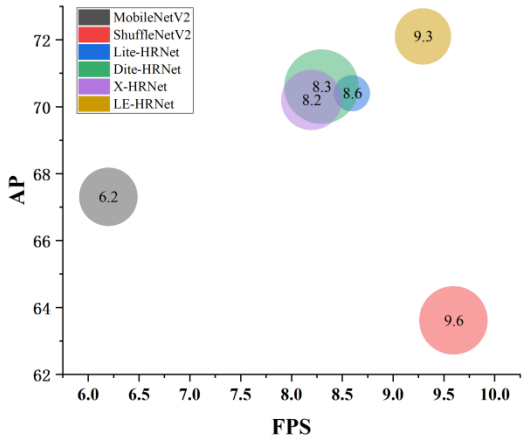


Fig. 7. Performance comparison of inference speed of different models.

The comparison of inference speed between some methods and LE-HNet was conducted in this paper. The testing of lightweight models emphasizes performance under limited resources, so the tests in this paper were conducted using only the CPU, specifically an Intel(R) Xeon(R) Silver 4310. The tests were carried out under consistent experimental conditions, and the results are shown in Fig. 7. Compared to other models, LE-HRNet achieves an inference speed of 9.6 FPS while maintaining high performance, making it faster than Lite-HRNet, Dite-HRNet, and X-HRNet. Although other lightweight models, such as ShuffleNetV2, have slightly faster inference speeds, their accuracy is lower and they fail to accurately detect human key points. LE-HRNet offers a better balance, and the verification of its inference speed demonstrates that LE-HRNet is more suitable for edge computing platforms.

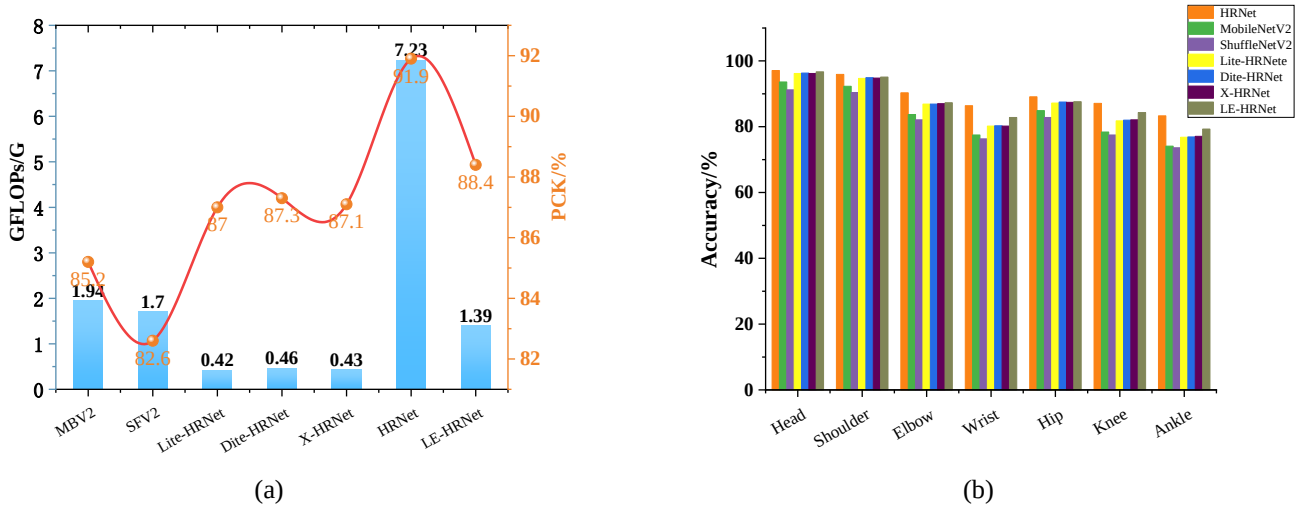
TABLE I. PERFORMANCE COMPARISON OF DIFFERENT ALGORITHMS ON THE COCO VALIDATION SET

Model	Params /10^6	GFlops/G	*AP*/%	AP^{50}/%	AP^{75}/%	AP^{M}/%	AP^{L}/%	*AR*/%
Hourglass	25.1	14.3	66.9	-	-	-	-	-
CPN	27.0	6.20	68.6	-	-	-	-	-
SimpleBaseline	34.0	8.90	70.4	88.6	78.3	67.1	77.2	76.3
HRNet	28.5	7.10	73.4	89.5	80.7	70.2	80.1	78.9
MobileNetV2	9.6	1.48	64.6	87.4	72.3	61.1	61.1	70.7
ShuffleNetV2	7.6	1.28	59.9	85.4	66.3	56.6	66.2	66.4
Lite-HRNet	1.8	0.31	67.2	88.0	75.0	64.3	73.1	73.3
Dite-HRNet	1.8	0.3	68.3	88.2	76.2	65.5	74.1	74.2
X-HRNet	2.1	0.3	67.4	87.5	75.4	64.5	73.3	73.5
LE-HRNet	5.4	1.51	69.3	88.6	77.2	66.1	74.3	74.6

Table II shows the comparison results with different human pose estimation algorithms on the MPII validation set. Compared to HRNet, LE-HRNet significantly reduces the number of parameters and computational load, with only a 3.6% decrease in accuracy. Compared to MobileNetV2 and ShuffleNetV2, LE-HRNet has lower parameters and an accuracy improvement of 3.0% and 5.6%, respectively. Compared to Lite-HRNet, Dite-HRNet, and X-HRNet, LE-HRNet improves accuracy by 1.7%, 1.1%, and 1.4%, respectively, with a slight increase in parameters and computational load. This demonstrates that LE-HRNet maintains model performance while ensuring low parameter count.

TABLE II. PERFORMANCE COMPARISON OF DIFFERENT ALGORITHMS ON THE MPII VALIDATION SET

Model	Params/10^6	GFLOPs/G	PCKh/%
MobileNetV2	9.6	1.9	85.4
ShuffleNetV2	7.6	1.7	82.8
Lite-HRNet	1.8	0.4	87.0
Dite-HRNet	1.8	0.4	87.6
X-HRNet	2.1	0.4	87.3
HRNet	28.5	7.6	92.3
LE-HRNet	5.4	1.44	88.7



(a) (b)

Fig. 8. Performance comparison of different algorithms on the MPII test set. (a) Comparison of different algorithms on GFLOPs and PCK; (b) Recognition accuracy of different algorithms at each keypoint.

Fig. 8 shows the performance comparison of different algorithms on the MPII test set. From the results in the figure, LE-HRNet has higher keypoint recognition accuracy compared to other lightweight algorithms. Particularly for some challenging keypoints such as Wrist, Knee, and Ankle, the accuracy improvement is more significant than for other keypoints. This is mainly due to the MCAM, which significantly enhances the detection capability for human keypoints.

We randomly selected a set of images from the COCO dataset for visual analysis. This set includes various scenarios such as single-person and multi-person scenes, as shown in Fig. 9. From the figure, it can be seen that LE-HRNet can accurately identify human keypoints in single-person scenarios.

Fig. 9. Visualization results on the COCO dataset.

In multi-person scenes, especially when there is overlap between body parts, LE-HRNet, with its strong feature extraction capabilities, can accurately infer the positions of the occluded keypoints by extracting multi-scale feature information and utilizing other visible keypoints. The visual analysis of different scenarios demonstrates that LE-HRNet maintains excellent detection performance in various complex scenes.

D. Ablation Experiment

To validate the effectiveness of the proposed method, we conducted ablation experiments on the COCO dataset for LEblock and CMFFM. The experimental results are shown in Table III. By replacing the residual block with LEblock, the model's parameter count is reduced by 76.4%, while the AP score only decreases by 4.5%, demonstrating that the lightweight module LEblock can significantly reduce the model's parameter count with minimal performance loss. Building on this, we inserted CMFFM for multi-scale feature fusion. The model's parameter count was further reduced, and the AP increased by 0.4%. This improvement is due to the optimization of the sampling process and the decomposition of the convolution process, which reduced the parameter count. Additionally, learning cross-resolution channel weight information effectively models channel features, and the spatial attention mechanism preserves more spatial detail features.

TABLE III. ABLATION EXPERIMENTS ON LEBLOCK AND CMFFM ON THE COCO DATASET

Model	LEblock	CMFFM	Params	AP/%
HRNet	×	×	28.5M	73.4
LE-HRNet	√	×	6.7M	68.9
	√	√	5.4M	69.3

To further validate the effectiveness of MCAM, we conducted ablation experiments on MCAM, with the results shown in Table IV. With the addition of MCAM, the parameter count increased by only 1.4M, while performance improved by 1.5%, demonstrating that MCAM effectively enhances the model's ability to detect human keypoints.

TABLE IV. ABLATION EXPERIMENTS ON MCAM

Model	Params	AP/%
+ELblock	6.7M	68.9
+ELblock (No MCAM)	5.3M	67.4

E. Discussion

To enable human pose estimation on mobile devices or edge computing devices, we propose a series of methods to streamline the high-resolution network. High-resolution networks are widely used in scenarios such as human pose estimation and semantic segmentation due to their high recognition accuracy. However, the high parameter count and computational complexity of these models make it difficult to deploy them on devices with limited computational resources. To address this, we propose replacing the standard 3×3 convolution with a Ghost module to reduce computational load, and we further optimize the upsampling and downsampling processes to improve computational efficiency. Additionally, to maintain model performance while reducing computation, we introduce a multi-scale coordinate attention mechanism that effectively minimizes performance loss due to lightweighting. Through this series of methods, we have successfully streamlined the high-resolution network and achieved favorable inference speed on low-power devices.

IV. Conclusion and Future Work

To address the issues of large parameter count and high computational complexity in high-resolution network models, we propose a lightweight and efficient high-resolution network module. We use the Ghost module to replace 3×3 convolution to reduce the parameter count and computational load of the model. Simultaneously, to minimize the loss of feature information during the lightweight process and ensure model performance, we designed a Multi-Scale Coordinate Attention Mechanism. This mechanism effectively enhances the detection of human keypoints by integrating multi-scale feature information and coordinate positional information without compromising performance. Finally, we optimized the multi-scale feature fusion stage, modeling both channel and spatial features while reducing the parameter count, further enhancing the model's performance. Experiments on multiple datasets validated the effectiveness of our proposed method.

In future work, we will deploy LE-HRNet on mobile devices and apply it in physical education, such as high jump, long jump, and swimming. By using LE-HRNet to identify key points of students' movements and calculate similarity with standard actions, we will be able to score students' movements and provide improvements for non-standard actions, which will aid in students' sports training.

Acknowledgment

This study was supported by Shandong Provincial Undergraduate Teaching Reform Project (Grant Number: Z2021450), Shandong Provincial Natural Science Foundation of P.R. China (Grant Number: ZR2020QF069), National College Students' Innovation and Entrepreneurship Training Program (Grant Number: 202310433069), and Shandong University of Technology Postgraduate Teaching Reform Project (Grant Number: 4053222063).

References

[1] L. Song, G. Yu, J. Yuan, and Z. Liu, “Human pose estimation and its application to action recognition: A survey,” J. Vis. Commun. Image Represent., vol. 76, p. 103055, 2021.

[2] C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, N. Kehtarnavaz, and M. Shah, “Deep learning-based human pose estimation: A survey,” ACM Comput. Surv., vol. 56, no. 1, pp. 1–37, 2023.

[3] P. F. Felzenszwalb and D. P. Huttenlocher, “Pictorial structures for object recognition,” Int. J. Comput. Vis., vol. 61, pp. 55–79, 2005.

[4] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[5] A. Toshev and C. Szegedy, “Deeppose: Human pose estimation via deep neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014, pp. 1653–1660.

[6] A. Newell, K. Yang, and J. Deng, “Stacked hourglass networks for human pose estimation,” in Proc. Comput. Vis. ECCV 2016: 14th Eur. Conf., Amsterdam, The Netherlands, Oct. 11-14, 2016, Part VIII, Springer, 2016, pp. 483–499.

[7] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, “Convolutional pose machines,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 4724–4732.

[8] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, “Realtime multi-person 2D pose estimation using part affinity fields,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 7291–7299.

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

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

[11] K. Sun, B. Xiao, D. Liu, and J. Wang, “Deep high-resolution representation learning for human pose estimation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 5693–5703.

[12] 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. 10440–10450, 2021.

[13] Q. Li, Z. Zhang, F. Xiao, F. Zhang, and B. Bhanu, “Dite-HRNet: Dynamic lightweight high-resolution network for human pose estimation,” arXiv preprint arXiv:2204.10762, 2022.

[14] Y. Zhou, X. Wang, X. Xu, L. Zhao, and J. Song, “X-HRNet: Towards lightweight human pose estimation with spatially unidimensional self-attention,” in 2022 IEEE International Conference on Multimedia and Expo (ICME), pp. 01–06, 2022, IEEE.

[15] 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.

[16] 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.

[17] 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.

[18] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, “ECA-Net: Efficient channel attention for deep convolutional neural networks,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 11534–11542, 2020.

[19] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in Proceedings of the European conference on computer vision (ECCV), pp. 3–19, 2018.

[20] Q. Hou, D. Zhou, and J. Feng, “Coordinate attention for efficient mobile network design,” in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 13713–13722, 2021.

[21] 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 Computer Vision--ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pp. 740–755, 2014, Springer.

[22] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, “2D human pose estimation: New benchmark and state of the art analysis,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3686–3693, 2014.

[23] 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.

[24] 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.