Lightweight Human Pose Estimation with Attention Mechanism

***Abstract*—In order to address the common issues in current human pose estimation network models, such as insensitivity to local information and inaccurate prediction of keypoint locations, various approaches have been proposed. To improve the accuracy of human pose estimation, a method is proposed which combines the CBMA attention mechanism with a lightweight high resolution network (Lite-HRNet). The method proposed in this paper was evaluated on the human keypoint detection datasets COCO and MPII, and compared with the current mainstream methods. The experimental results show that the proposed method effectively improve the accuracy of the model while ensuring minimal loss in computation time.**

Keywords*—*human pose estimation, attention mechanism, lightweight

I. Introduction

Human pose estimation is an important research direction of computer vision. In recent years, with the development of deep learning technology, related research results are widely used in the fields of behavior detection, video surveillance, virtual reality and so on. However, the development of human pose estimation technology is still restricted by the bottleneck problems such as lack of global feature relationship, scale difference of target and occlusion of key points in practical application.

Human pose estimation approaches based on deep learning can be divided into top-down and bottom-up methods. The top-down method first performs human object detection on the input image to obtain human objects with bounding boxes. Then, the bounding box is cropped to the size of a single human body, and feature extraction is performed using the pose estimation network to obtain the coordinates of each keypoint of the human body. In 2016, Wei et al. [1] designed the convolutional pose machine network, which uses convolutional layers to express texture information and spatial information, and designed a multi-stage structure to improve the detection performance of single keypoints. In 2017, Fang et al. [2] designed the regional multi-person pose estimation network, focusing on the problems of detection frame positioning error and repeated detection in the top-down method of target detection algorithms. The human body bounding box is optimized by the spatial transformation network, which overcomes the inflfluence of the target detection algorithm error on the subsequent keypoint detection task. In 2021, Yu et al. [3] designed Lite-HRNet, which integrates ShufflfleNet into a high-resolution network and reduces the computational complexity while improving performance. Network models such as MobileNet and ShufflfflffleNet are designed to make the model smaller and faster by employing a more effiffifficient network structure rather than compressing or migrating a large trained model.

The bottom-up method fifirst performs global keypoint detection on the input image to obtain all keypoints in the image. Then, using the positional relationships of human joints, the joint points are combined into multiple groups of independent human keypoints using a clustering algorithm. In 2017, Cao et al. [4] proposed Openpose and designed a classic keypoint clustering algorithm called part affiffiffinity fifields that can simultaneously encode the position and direction of joint points to balance keypoint detection speed and accuracy. In 2022 Li et al [5] proposes an improved method for human pose estimation using Lite HRNet and a coordinate attention mechanism. Lite HRNet is a computer vision model that balances computational complexity with high precision by using a channel attention mechanism instead of complex convolution.

Recent researches also suggest that representations produced by CNNs can be further strengthened by integrating attention learning mechanisms into the network. By considering the dependency relationship between feature channels, importance of features can be captured with automatic learning process. Therefore, features that contribute more will be enhanced while those with little importance will be suppressed.

For solving the key problems of human pose estimation technology, this paper proposes the High-Resolution Network, HRNet as a benchmark model, introduces attention mechanism from channel, space and scale, and integrates it into the benchmark model. On this basis, data enhancement is used to improve the performance of the model, so as to achieve the accurate extraction of the target key information.

II. RELATED WORK

A. *Lightweight High-Resolution Network*

High-resolution Net (HRNet) is fifirst proposed by Sun *et al.* for human pose estimation [6]. It maintains high-resolution representations throughout the network by connecting high to-low resolution convolutions in parallel, where feature maps of low resolution are added to that of high resolution gradally. There have been a number of mature studies on Network Lightweight Research. Most of the current pose estimation tasks use high-resolution networks as the backbone network, and the pose estimation networks proposed in the past two years such as HigherHRNet [11] and DEKR [12] have also been designed and improved based on this network.The Mobilenet series [7-9] mainly propose deep separable convolution and inverse residual structure that can reduce a large number of parameters and increase the computational speed, which is the first choice of many lightweight models. Shufflenet [10] group the input features and then channel-mixing wash them to reduce the computational effort while ensuring that the information in each group flow.

Currently, human pose estimation is still mainly done through convolutional networks to predict each key point and many researches introduced transformer[13] into pose estimation and even gradually replace convolutional networks, reducing the number of model parameters compared to convolutional networks, but there is still the problem of low computational efficiency. In addition, some studies aim at lightweight self-attention methods. A representative one is the window-attention of Swin transformer [14], in which the whole feature map is divided into multiple small windows of the same size, and each window completes the calculation of self-attentionindependently, thus improving the computational efficiency of the model. The above method only achieves model lightweighting, but the model accuracy is reduced. To address this issue, I propose adding an attention mechanism to the model, which can improve the model accuracy.

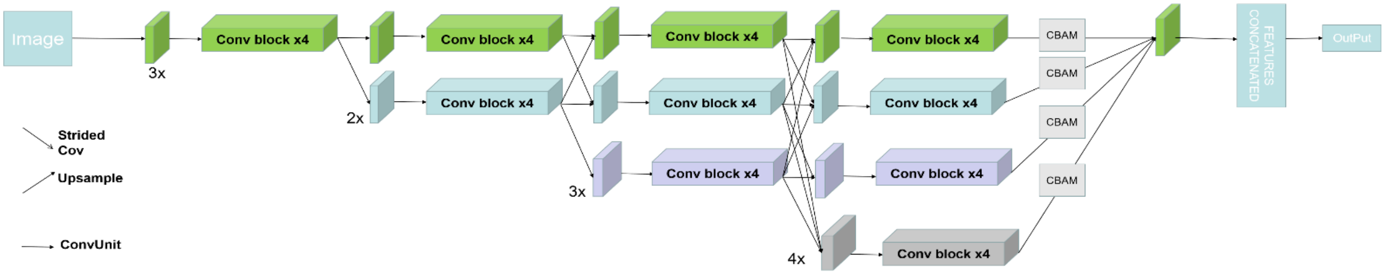

Fig. 1.The block diagram of the modified light-HRNet architecture is shown below. The output of each upsampling layer is extracted and passed through a CBAM module (as indicated by the gray box). The results of each attention block are concatenated and input into the fully connected layer.

B. *Attention Mechanisms*

Attention mechanisms can be roughly divided into two categories: strong attention and soft attention mechanisms. Because strong attention is a random prediction that emphasizes dynamic changes, although its performance is good, its application is very limited because of its non-differentiable nature. On the contrary, soft attention is differentiable everywhere, that is, it can be obtained by neural network training based on the gradient descent method, so its application is relatively wide. A soft attention mechanism is divided according to the different dimensions of attention. The current mainstream attention mechanism can be divided into the following three types: channel attention, spatial attention, and self-attention. For channel attention, the purpose is to model the correlation between different channels (feature maps), automatically obtain the importance of each feature channel through network learning, and finally assign different weights to each channel. Weight coefficients are used to strengthen important features and suppress non-important features.

III. Lightweight Model Structure

This section focuses on the construction process of the pose estimation model. The construction process of the pose estimation model consists of four steps: dataset selection, training process, testing process, and evaluation process.

A. *Datasets*

COCO [15] has over 200K images and 250K person instances with 17 keypoints. Our models are trained on train2017 dataset (includes 57K images and 150K person instances) and validated on val2017 (includes 5K images) and test-dev2017 (includes 20K images).The MPII Human Pose dataset [16] contains around 25K images with full-body pose annotations taken from real world activities. There are over 40K person instances, split 12K instances for testing, and others for training.

B. *Light-HRNet with CBAM Model*

I found that a lightweight unit and a conditionally weighted channel were introduced to replace the expensive pointwise (1x1) convolution in the shuffle block of HRNet. Although this achieved lightweight design, there was a loss of accuracy. To address this issue, I added CBAM modules before the fully connected layer in each resolution of light-HRNet to pay more attention to the detailed information of key points, improving the accuracy while keeping the complexity relatively stable.

C. *Convolutional Block Attention Modules*

To inculcate attention into our work, we used the convolutional block attention module proposed by Sanghyun Woo et al. [17]. We use this because the module can be used as an additional plugin to our skip connections and it is end to end trainable.The module can be divided into two parts which are spatialand channel attention submodules. Given an intermediate feature map $F \in IR^{C\times H\times W}$ as input, CBAM in Fig .2. sequentiallyinfers a 1D channel attention map $M_c \in IR^{C\times 1\times 1}$ and a 2D spatial attention map $M_S \in IR^{1\times H\times W}$. The overall transformation performed by the module can be summarized as:

$$F' = M_c(F) \otimes F \qquad (1)$$

$$F'' = M_C(F') \otimes F' \qquad (2)$$

where ⊗ denotes element-wise multiplication. F'' is the final refined output after being processed by the attention module.

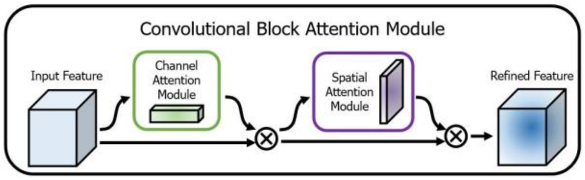

Fig. 2. **The overview of CBAM** The module has two sequential sub-modules:channel and spatial. The intermediate feature map is adaptively refined throughour module (CBAM) at every convolutional block of deep networks.

D. *Implementation details*

For both data sets, we use the following strategy to split the data conventionally for facilitating our comparison with the latest methods: 80% as the training set and 20% as the test set. In order to evaluate the results of our model for scene classification, overall accuracy (OA) and confusion matrix are used for the following experiments. **Traning** Furthermore, all models are trained using stochastic gradient descent (SGD) with a batch size of 256 for 150 epochs. The initial learning rate is 0.0025, which is divided by a factor of 10 after 40 and 80 epochs.All experiments are conducted on a PC with Ubuntu 20.04, PyTorch deep learning framework, CPU i7-11700K and GPU GTX-2080Ti. **Testing.** We estimate the heat maps via a post-gaussian filter and average the original and flipped images' predicted heat maps. A quarter offset in the direction from the highest response to the second-highest response is applied to obtain each keypoint location. **Evaluation.** We adopt the OKS-based mAP metric on COCO, where OKS (Object Keypoint Similarity) defines the similarity between different human poses. We report standard average precision and recall scores: AP (the mean of AP scores at 10 positions, OKS = 0.50, 0.55, . . . , 0.90, 0.95), AP50 (AP at OKS = 0.50), AP75, AR and AR50. For MPII, we use the standard metric to evaluate the PCKH @0.5 (head-normalized probability of correct keypoint) performance.

IV. Experiment results

We evaluate the performance of CBAM-LiteHRNet and other state of-the-art models on COCO and MPII datasets models in the evaluation. Table 1 and Table 2 show the overall accuracy of different. It can be observed that the proposed CBAM-LiteHRNet model has achieved the best performance compared with other models on both datasets. With the addition of CBAM module, the proposed CBAM-LiteHRNet has outperformed state-of-the-art methods,The employ of SE module has remarkably improved CBAM-HRNet on overall accuracy by 0.1% on two data sets, which exhibits the effectiveness of thiscombination.

A. *Analysis of the Results*

Table 1 shows the experimental results on the COCO val2017 dataset. The results show that when the input resolution is 256 × 192, the AP value of CMAB-LiteHRNet is 67.0, which is almost 0.1% better than the baseline network Lite-HRNet, but the number of network parameters is not as high as that of the baseline network. By increasing the input image scale and the number of input channels, the detection accuracy can be further improved. When the input resolution is 384×288 and the number of channels is 48, the best performance is achieved, yielding an AP of 69.9. Moreover, the number of parameters is still lower than the baseline network with 32 channels. Compared with other classical pose estimation methods, it performs better with respect to parameter quantity and average accuracy.

TABLE I. Results of each model on COCO val2017

Model	***Input size***	***AP***	***AP^{50}***	***AP^{75}***	***AP^{L}***	***AR***
Large networks						
HRNetV1-W32[6]	384 × 288	74.9	*92.5*	*82.8*	*80.9*	*80.1*
HRNetV1-W48[6]	384 × 288	*75.5*	*92.5*	*83.3*	*81.5*	*80.5*
DARK[19]	128 × 96	*76.2*	*92.5*	*83.6*	*82.4*	*81.1*
SimpleBaseline[20]	384 × 288	*73.7*	*91.9*	*81.1*	*80.0*	*79.0*
Small networks						
Lite-HRNet-18	384 × 288	*66.9*	*89.4*	*74.4*	*72.2*	*72.6*
Lite-HRNet-30	384 × 288	*69.7*	*90.7*	*77.5*	*75.0*	*75.4*
MobileNetV2 1× [18]	256 × 192	*64.6*	*87.4*	*72.3*	*71.2*	*70.7*
OurNet	256 × 192	***67.0***	***90.0***	***74.6***	***74.4***	***72.1***
OurNet	384 × 288	***69.9***	***91.0***	***77.3***	***77.0***	***75.3***

TABLE II. Experimental results on MPII validation set

Model	***Type***	***Head***	***Wrist***	***Hip***	***Knee***	***Ankle***	***Mean***
SimpleBaseline[20]	HM	97	85	89.2	85.3	81.3	89.6
HRNetV1-W32[6]	HM	96.9	85.8	88.7	86.6	82.6	90.1
TokenPose [21]	HM	97.1	95.9	89.3	87.1	82.5	90.2
Lite-HRNet	HM	91	86.2	88.4	83.4	88.2	87
OurNet	HM	93	86.4	88.5	84.1	88.4	87.1

The CBAM-LiteHRNet algorithm proposed in this paper is compared with other advanced pose estimation algorithms proposed in recent years. Table Ⅱ shows the results on the MPII validation set. The CBAM-LiteHRNet algorithm uses approximately a quarter of the parameters used by the baseline network Lite-HRNet, but achieves a 0.5 percentage point improvement in accuracy. Compared with other recent attitude estimation methods, our network is better in parameter quantity and accuracy.

V. CONCLUSION

The CBAM-HRNet proposed in this paper is an improved version of the high-resolution network. The Lite-HRNet was combined with the CBAM module to create a new lightweight module to replace the basic module in the feature extraction network while reducing the number of network parameters and retaining accuracy of the network model.

References

[1] Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 4724–4732 (2016). *(references)*

[2] Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: Rmpe: Regional multi-person pose estimation. In: Proceedings of the IEEE international conference on computer vision, pp. 2334–2343 (2017)

[3] Yu, C., Xiao, B., Gao, C., Yuan, L., Zhang, L., Sang, N., Wang, J.: Lite-hrnet: A lightweight high-resolution network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10440–10450 (2021).

[4] Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affiffiffinity fifields. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7291–7299 (2017).

[5] Li R, Huang H, Zheng Y. Human pose estimation based on lite HRNet with coordinate attention[C]//2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP). IEEE, 2022: 1166-1170.

[6] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5693–5703.

[7] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.

[8] Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4510-4520

[9] Howard A, Sandler M, Chu G, et al. Searching for mobilenetv3[C]//Proceed-ings of the IEEE/CVFinternational co-nference on computer vision. 2019: 1314-1324

[10] Zhang X, Zhou X, Lin M, et al. Shufflenet: An extremely efficient convolutional neural network for mobile devices[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6848-6856

[11] Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T.S., Zhang, L.: 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)

[12] Geng, Z., Sun, K., Xiao, B., Zhang, Z., Wang, J.: Bottom-up human pose estimation via disentangled keypoint regression. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14676–14686 (2021)

[13] Vaswani A, Shazeer N, Parmar N, et al.Attention is all you need[J].Advances in neural information processing systems, 2017, 30.

[14] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021:10012-10022.

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays,Pietro Perona, Deva Ramanan, Piotr Doll´ar, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In Proc. European Conference on Computer Vision (ECCV), 2014. 5

[16] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmarkand state of the art analysis. In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3686–3693, 2014. 5

[17] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In-So Kweon. Cbam: Convolutional block attention module. In ECCV, 2018.

[18] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Inverted residuals and linear bottlenecks: Mobile networks for classification. Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1801, 2018. 2, 5, 8

[19] Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 7093–7102, 2020. 5, 6

[20] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In Proc. European Conference on Computer Vision (ECCV), pages 466–481, 2018. 5, 6

[21] Li, Y., Zhang, S., Wang, Z., Yang, S., Yang, W., Xia, S.T., Zhou, E.: Tokenpose: Learning keypoint tokens for human pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11313–11322 (2021)