High-Resolution with Global Context Network for Human Pose Estimation

1st Kehao Wang
School of Information Engineering
Wuhan University of Technology
Wuhan, China
kehao.wang@whut.edu.cn

2nd Chenglin Li
School of Information Engineering
Wuhan University of Technology
Wuhan, China
chenglin.li@whut.edu.cn

3rd Ruiqi Ren
School of Information Engineering
Wuhan University of Technology
Wuhan, China
ruiqi.ren@whut.edu.cn

***Abstract*—Despite significant improvement in human pose estimation research, most top-performance methods are challenging to deploy in practical applications because of their complex architecture and high computational costs. Although the lightweight human pose estimation approach requires less processing and may be deployed on devices with low resources, such as mobile phones or robots, its network model performance is not exceptional. In this paper, we design the structure based on High-Resolution Network (HRNet), and propose a High-Resolution and Global Context Network (HRGCNet) based on the attention mechanism. Our approach redesigns the bottleneck block according to the attention mechanism of the Global Context Network (GCNet). By combining lightweight and high-performance GC blocks with bottleneck blocks, HRGCNet adds global context features at each location in the high-resolution subnet. The resulting high-resolution representation contains richer feature information. Our experiments on the COCO train2017 dataset show the efficiency of our method. Compared to HRNet with state-of-the-art performance, HRGCNet achieves higher accuracy, and the AP score improves by 2.0 percentage points with similar model size (#Params) and computational complexity (FLOPs). On the COCO test-dev set, HRGCNet has an AP score of 78.3, which is better than most current methods with good performance.**

***Index Terms*—Human pose estimation, HRNet, GCNet, depth-wise separable convolution**

I. Introduction

Human pose estimation refers to labeling the position of joint human points in a picture or video and optimally connecting them. It has become a popular research direction of computer vision, with applications in motion recognition [1], human-computer interaction [2], and auto-drive [3]. Due to the diversity of human pose, the complexity of the surroundings, and the ambiguity of the perspective, human pose estimation faces great obstacles.

Recently, the deep convolution neural network has significantly improved human pose estimation. For example, HRNet [4] learns to get reliable high-resolution heatmaps by concurrently connecting multi-resolution subnetworks and performing multi-scale fusion repeatedly. Using multi-resolution supervised training and multi-resolution aggregation inference, HigherHRNet [5] can solve the scale change problem in bottom-up pose estimation. In addition, with the significant improvement of attention mechanisms in target detection, image classification, etc., it has achieved good results in human pose estimation. For example, Li [6] proposed a regression-based pose recognition method using a cascade transformer, which uses the encoder-decoder structure in transformers to perform regression-based pedestrian and keypoint detection, revealing the recursive self-attention mechanism in the transformer. Xu [7] first proposed a simple baseline model ViTPose based on vision transformer structure. The model only uses vision transformer structure as encoder and a few deconvolution layers as decoder, which can achieve good performance in human pose estimation tasks.

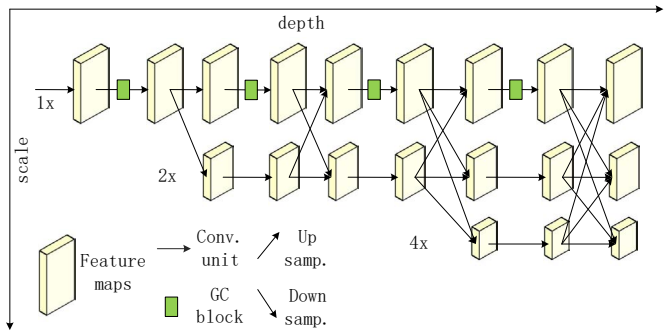


Fig. 1. The architecture of the proposed HRGCNet

However, most advanced approaches have complex architectures with numerous parameters and floating-point computations. Although these methods achieve the best performance, they require much memory because of these models' large number of parameters. On the other hand, due to numerous floating-point operations, they are time-consuming during the training phase. In addition, to deploy the trained network to devices with limited resources, such as mobile phones or robots, there is an increasing demand for a network of human pose estimation with fewer parameters, less computation, and high accuracy. For example, Osokin [8] proposed Lightweight OpenPose based on OpenPose. Compared with the second-order OpenPose, the parameter amount is only 15%, but the performance is similar. Yu [9] proposed a Lightweight High-Resolution Network (Lite-HRNet) and applied the efficient ShuffleNet block in ShuffleNet to HRNet. Zhang [10] presented a Lightweight Pose Network (LPN) and redesigned the network model based on SimpleBaseline architecture with

deep convolution and attention mechanism. These networks have the advantages of small model size and less floating-point computation. They can be deployed in practical applications, but their accuracy is not comparable to the most advanced methods.

In this work, we focus on how to improve the performance of the human pose estimation network model without increasing the computation cost. The contributions of this paper are as follows.

- Based on a new attention mechanism, we combine GC block with bottleneck block and apply it to HRNet. We introduce an attention mechanism, which adds global context features to each location of the high-resolution feature map and obtains HRGCnet.
- We add depthwise separable convolutions to ensure the capture of spatial information and the fusion of information across channels.
- Experiments show that our algorithm has more advantages, the AP score of HRGCNet can reach 78.3 on the COCO test development set, which is significantly better than the high-resolution network HRFormer with SOTA performance.

II. Related Work

A. Human pose estimation

The deep learning model shows better performance than the traditional methods in human pose estimation. For example, they use deep convolution neural networks (DCNNs) to anticipate keypoints in the human body [11], [12]. He [13] uses a feedforward connection approach to ensure that the back-propagation gradient is maintained, allowing the network layers to continue deeper. Simonyan [14] extracts the network with Visual Geometry Group Network (VGGNet) as the feature and uses a multi-scale feature cascade to compensate for the loss of feature information during the pooling procedure.

Bottom-up approach. The bottom-up method groups all body joints into corresponding individuals in one image. Xiao [15] added some deconvolution layers to ResNet in order to create a simple and effective structure for generating heatmaps for high-resolution representation while also reducing the algorithm's complexity. Cao [16] presented Openpose and used a nonparametric representation method of Part Affinity Fields (PAFs) to learn to associate the target object with the body keypoints, reducing the calculation time. The multi-person pose estimation methods can be divided into top-down and bottom-up.

Top-down approach. The top-down method first detects everyone in the image, then estimates everyone's pose using the single-person pose estimation method. G-RMI [17] estimates pose by locating critical body points based on activation heatmap and replaces box-level scores with confidence score estimates based on critical points to avoid repeated pose detection. Mask R-CNN [18] estimates the pose of the human body by generating one-hot masks, but it cannot handle problems such as obscured key points, invisible keypoints, and crowded backgrounds.

However, most existing methods recover high-resolution representations from low-resolution representations. Sun [4] proposed HRNet to learn reliable high-resolution heatmaps by concurrently connecting multi-resolution subnetworks and performing repetitive multi-scale fusion. In addition to target detection and semantics segmentation, a high-resolution net demonstrates high test accuracy. As a consequence, we use HRNet as our benchmark network and improve its human pose estimation performance.

B. Attention mechanism

In recent years, attention mechanism plays an increasingly important role in computer vision, such as semantic segmentation [19], face recognition [20], motion recognition [21], and so on. The attention mechanism originates from the study of human vision. When people look at things, they selectively focus on the part of the whole information while ignoring other visible information. Based on this attention mechanism, there are two main contents: one is to determine which part of the input information needs more attention, and the other is to extract the features of critical parts to get important information.

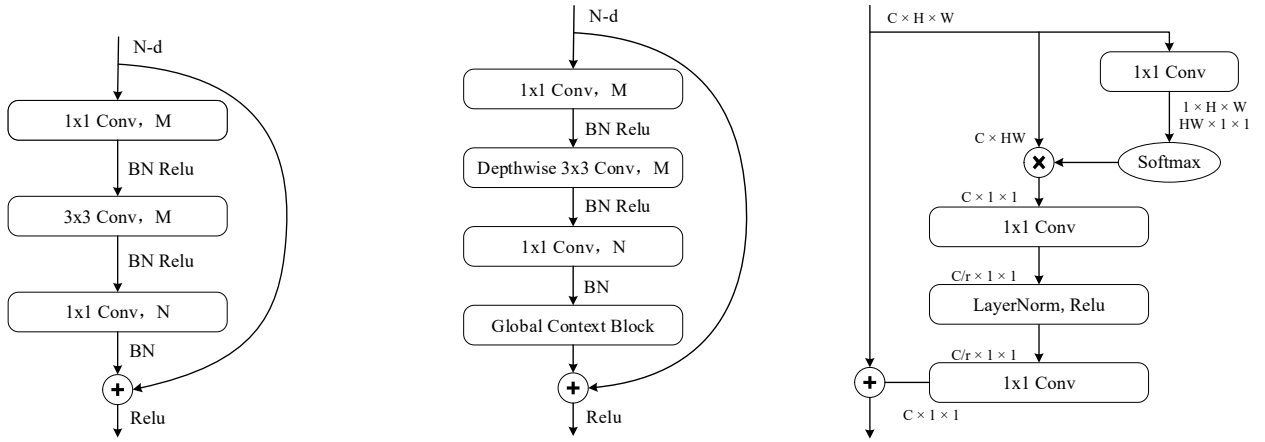
Chu [22] applied the attention mechanism to the human pose estimation model for the first time and proposed a method to integrate convolutional neural networks with the context attention mechanism into an end-to-end framework. Wang [23] proposed a Non-Local Network (NLNet) that uses self-attention to model pairwise relationships at the pixel level to capture long-term dependencies. It improves the performance of human pose estimation, but non-local operations in NLNet learn query-independent attention maps for each query location, which results in computational overhead. Hu [24] presented a new architecture unit called the Squeeze-Excitation (SE) module, which scales different channels to rebalance channel dependency and the global context. These blocks can be stacked together to form a SENet architecture, which can significantly improve model performance with a slight increase in computing costs.

Through rigorous empirical analysis, Cao [25] found that the global context of NLNet modeling is almost the same for different query locations in the image. They designed a better instantiation, called GC block, and then built a global context network, which can effectively model the global context by adding fusion. As a consequence, it is applied to bottleneck blocks of the high-resolution network HRNet, which can increase network capacity without putting too much computational strain on the system and improve model performance by increasing global context characteristics for each location in the high-resolution signature graph.

III. Approach

A. HRNet

Previously, the high-resolution network for mainstream keypoint detection consisted of a series of connected high-to-low-resolution subnets, which output high-resolution representations through an up-sampling operation. Such networks do



(a) Bottleneck Block (b) Lightweight Bottleneck with GC Block (c) Global Context Block

Fig. 2. Architecture of the main blocks. (a) Standard Bottleneck Block in ResNet. (b) Lightweight Bottleneck with GC Block. The redesigned Lightweight Bottleneck Block after two modifications. Note that M and N in these blocks denote the number of output channels of a convolutional layer. (c) Global Context Block, which is lightweight and can effectively model long-range dependency.

not fully compensate for the loss of information in spatial resolution due to up-sampling operations, resulting in the low spatial sensitivity of the high-resolution representations of the final output. It is primarily limited by the resolution of the semantically expressive representation. There are also high-resolution networks that connect subnets of different resolutions in parallel, but there is no multi-scale information exchange between subnets. And the high-resolution representation of the final output is only obtained from the original high-resolution representation after a few convolution operations, resulting in the high-resolution representation of the network's final output providing only low-level semantic representation.

HRNet is different from these two types of high-resolution networks. Throughout the process, HRNet maintains high-resolution representations. It keeps high-resolution representations throughout the network by connecting subnetworks from high to low at the same time. It also performs multi-scale fusion consistently, resulting in more accurate heatmaps.

HRNet starts with two 3 × 3 convolutions, reduces the resolution to a quarter of the original, then regresses the heatmaps at that resolution. There are four phases to the high-resolution subnetwork. A high-resolution subnetwork is part of the initial phase. The subsequent three phases gradually reduce the high-resolution subnetwork's resolution until it reaches the low-resolution subnetwork. While the number of channels doubles, the resolution falls. Each phase connects subnetworks with varying resolutions in parallel and repeatedly performs information fusion at various scales. Finally, the high-resolution output heatmap predicts the location of significant locations.

HRNet has two distinct advantages over other networks for pose estimation: (i) Instead of serial connections to high-to-low-resolution subnets, as most existing systems do, parallel connections are used, allowing the high resolution to be preserved rather than restored through a low-to-high procedure. (ii) Lower and higher representations are combined in most existing fusion schemes. Instead, HRNet uses recurrent multi-scale fusion, which improves high-resolution representations by combining low-resolution representations of the same depth and related levels, allowing high-resolution representations to be used for pose assessment.

B. Global Context Block

Cao [25] proposed a new global context modeling framework instance called global context block. It has the advantages of a simplified non-local block, which can effectively model long-distance dependencies, and squeeze-excitation block, which can perform lightweight computing. They use NLNet to abstract the general framework for global context modeling: (a) Global attention pool, which gets attention weights using 1x1 convolution W_k and Softmax functions, then uses attention pool to gain global context features. (b) Feature transformation through 1x1 convolution W_v. (c) Feature aggregation, which aggregates global context features into features at each place by using addition. The general framework can be written as follows

$$z_i = F\left(x_i, \delta\left(\sum\nolimits_{j=1}^{N_p} \alpha_j x_j\right)\right) \tag{1}$$

where N_p is the number of positions in the feature map, $x = \{x_i\}_{i=1}^{N_p}$ denotes the feature map of one input instance, $\sum_j \alpha_j x_j$ denotes context modeling by grouping the characteristics of all locations together by weighted averaging of weighting α_j to get a global attention pool in a simplified NL (SNL) block, $\delta(\cdot)$ denotes a feature transformation used to capture channel correlation, $F(\cdot,\cdot)$ is a fusion function representing features that aggregate global context features into each location.

The author finds that SE block is actually an example of this context modeling framework. For SE blocks, $\alpha_j = \frac{1}{N_p}$, $\delta(\cdot)$ is a 1x1 convolution, a ReLU, a 1x1 convolution, and a sigmoid function, $F(\cdot,\cdot)$ multiplies the corresponding channel

broadcasts. Finally, the detailed architecture of the global context (GC) block is shown in Figure 2 (c), and its formula is defined as

$$z_i = x_i + W_{v2} ReLU\left(LN\left(W_{v1}\sum_{j=1}^{N_p}\alpha_j x_j\right)\right) \quad (2)$$

where W_{v1}, W_{v1} denote linear transform matrices, and α_j is the weight of the global attention pool, it is defined as

$$\alpha_j = \frac{e^{W_k x_j}}{\sum_{m=1}^{N_p} e^{W_k x_m}} \quad (3)$$

where W_k denotes a linear transform matrice, and $\delta(\cdot)$ is the bottleneck transform, it's defined as

$$\delta(\cdot) = W_{v2} ReLU\left(LN\left(W_{v1}\left(\cdot\right)\right)\right) \quad (4)$$

Figure 2 (c) shows the details of the GC block. Although SE block is an instantiation of the Global Context Modeling Framework, there is a difference. SE block strengthens important channels, weakens less important ones, and manipulates attention on channels. In contrast, GC blocks add a global context to each position rather than manipulating attention on positions.

C. HRNet with Global Context Block

HRNet can output more reliable high-resolution representations than other human pose estimates, allowing it to predict heatmaps more accurately in space. However, the architecture of HRNet is complex, and its high memory and time-consuming requirements make it more challenging to improve the model's performance without increasing the computing costs. Cao [25] noticed that the thermograms of attention for different query points were almost identical, indicating that although non-local solves the problem of long-distance dependence through self-attention, experiments show that attention maps are not dependent on the location of the query. That is, they have not learned attention.

GCNet combines the benefits of both NLN and SE blocks, allowing them to successfully simulate long-distance dependency, obtain global context information, and be used for lightweight computing. Therefore, we combine lightweight, high-performance GC blocks with bottleneck blocks. As shown in Figure 2 (a), a bottleneck block consists of three convolutional layers and a shortcut connection. Figure 2 (b) and (c) show the architecture of bottleneck blocks with GC blocks and the details of GC blocks, respectively.

As shown in Figure 1, the horizontal and vertical directions correspond to the network's depth and the feature map's scale, respectively. Like HRNet, our HRGCNet consists of parallel high-resolution to low-resolution subnets and performs repetitive information exchange between multi-resolution subnets. However, the difference is that we add lightweight GC blocks between high-resolution subnets. We apply a Bottleneck block with a lightweight GC block to a high-resolution subnet of HRNet to increase network capacity without increasing computation costs. In addition, by adding global context features to each location of the high-resolution signature map, the high-resolution representation of the final output contains richer feature information, which improves the performance of the model.

IV. Experiments

A. Setting

COCO keypoint detection. COCO datasets contain more than 200K images and 250K instances marked with 17 key points. The COCO train2017 dataset, which contains 57K pictures and 150K personal instances, was used to train our model. The val2017 and test-dev2017 episodes, each with 5K and 20K photos, were used to evaluate our technique. A statistic for evaluating anything. Object Keypoint Similarity (OKS) is the standard evaluation metric

$$OKS = \frac{\sum_i \exp\left(-d_i^2/2S^2k_i^2\right)\sigma\left(v_i > 0\right)}{\sum_i \sigma\left(v_i > 0\right)} \quad (5)$$

where d_i is the Euclidean distance between the detected key point and its corresponding ground true value, v_i is the visibility marker off the ground true value, S is the object proportion, and k_i is the constant of each key point controlling the attenuation. We used standard accuracy and recall rates to represent test results: AP (average detection accuracy at $OKS = 0.50, 0.55, 0.60, ..., 0.90, 0.95$), AP^{50} (detection accuracy at $OKS = 0.50$), AP^{75}, AP^M for medium-scale targets, AP^L for large-scale targets, AR for average recall at $OKS = 0.50, 0.55, ..., 0.90, 0.95$.

Training. The images of the COCO dataset are adjusted to 256 × 192 or 388 × 284, and some images are randomly rotated [45°, 45°], scaled randomly [0.65, 1.35], or flipped to expand the additional data of the coco dataset. The primary learning rate was set at 1e-3 and decreased tenfold in the 170th and 200th epochs, respectively, until it finally ended in the 210th epoch.

Test. We detect people instances through a person detector and then predict keypoints, where the person detectors for both the validation set and the test set are SimpleBaseline. As a general practice, we obtain heatmaps by averaging the predicted heatmaps of the original and flipped images. Apply a quarter offset in the direction from the highest response to the second-highest response to obtain each critical point location.

B. Results

1) COCO validation: Table I shows our comparison with other advanced methods. At the same time, Figure 3 shows the comparisons of the AP scores, GLOPs, and params of HRGCNet on the COCO val set with these top-level performance methods mentioned in Table I. The input size of the model is 256 × 192, and the bubble size indicates the number of model parameters. According to the width of the high-resolution subnet and the input size of the image in the last three stages, we have four different versions of HRGCNet. Our small network, HRGCNet-W32, is trained with 256 × 192

TABLE I
COMPARISONS OF RESULTS ON COCO VALIDATION SET. #PARAMS AND FLOPS ARE CALCULATED ONLY FOR THE POSE ESTIMATION NETWORK.

Method	Backbone	Input size	#Params	FLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
CPN [26]	ResNet-50	256 × 192	27.0M	6.2G	68.6	-	-	-	-	-
CPN+OHKM [26]	ResNet-50	256 × 192	27.0M	6.2G	39.4	-	-	-	-	-
SimpleBaseline [15]	ResNet-50	256 × 192	34.0M	8.9G	70.4	88.6	78.3	67.1	77.2	76.3
SimpleBaseline [15]	ResNet-101	256 × 192	53.0M	12.4G	71.4	89.3	79.3	68.1	78.1	77.1
SimpleBaseline [15]	ResNet-152	256 × 192	68.6M	15.7G	72.0	89.3	79.8	68.7	81.9	77.8
HRNet-W32 [4]	HRNet-W32	256 × 192	28.5M	7.1G	74.4	90.5	81.9	70.8	81.0	79.8
HRNet-W48 [4]	HRNet-W48	256 × 192	63.6M	14.6G	75.1	90.6	82.2	71.5	81.8	80.4
TransPose-H-A6 [27]	HRNet-W48	256 × 192	17.5M	21.8G	75.8	-	-	-	-	80.8
TokenPose-L/D24 [28]	HRNet-W48	256 × 192	27.5M	11G	75.8	90.3	82.5	72.3	82.7	80.9
HRFormer-B [29]	HRFormer	256 × 192	43.2M	12.2G	75.6	90.8	82.8	71.7	82.6	80.8
SimpleBaseline [15]	ResNet-152	384 × 288	68.6M	35.6G	74.3	89.6	81.1	70.5	79.7	79.7
HRNet-W32 [4]	HRNet-W32	384 × 288	28.5M	16G	75.8	90.6	82.7	71.9	82.8	81.0
HRNet-W48 [4]	HRNet-W48	384 × 288	63.6M	32.9G	76.3	90.8	82.9	72.3	83.4	81.2
HRGCNet-W32(ours)	HRGCNet	256 × 192	29.6M	7.11G	76.6	93.6	84.6	73.9	80.7	79.3
HRGCNet-W48(ours)	HRGCNet	256 × 192	64.6M	14.6G	77.4	93.6	84.8	74.6	81.7	80.1
HRGCNet-W32(ours)	HRGCNet	384 × 288	29.6M	16.1G	78.0	93.6	84.8	75.0	82.6	80.5
HRGCNet-W48(ours)	HRGCNet	384 × 288	64.6M	32.9G	**78.4**	**93.6**	**85.8**	**75.3**	**83.5**	**81.3**

TABLE II
COMPARISONS ON THE COCO TEST-DEV SET. #PARAMS AND FLOPS ARE CALCULATED ONLY FOR THE POSE ESTIMATION NETWORK.

Method	Backbone	Input size	#Params	FLOPs	AP	AP_{50}	AP_{75}	AP_{M}	AP_{L}	AR
Mask-RCNN [18]	ResNet-50	-	-	-	63.1	87.3	68.7	57.8	71.4	-
G-RMI [17]	ResNet-101	353 × 257	42.6M	57.0G	64.9	85.5	71.3	62.3	70.0	69.7
SimpleBaseline [15]	ResNet-50	256 × 192	34.0M	8.9G	70.0	90.9	77.9	66.8	75.8	75.6
SimpleBaseline [15]	ResNet-152	256 × 192	68.6M	15.7G	71.6	91.2	80.1	68.7	77.2	77.3
CFN [30]	-	-	-	-	72.6	86.1	69.7	78.3	64.1	-
TransPose-H-A6 [27]	HRNet-W48	256 × 192	17.5M	21.8G	75.0	82.2	82.3	71.3	81.1	-
TokenPose-L/D24 [28]	HRNet-W48	256 × 192	29.8M	22.1G	75.9	92.3	83.4	72.2	82.1	80.8
SimpleBaseline [15]	ResNet-152	384 × 288	68.6M	35.6G	73.7	91.9	81.1	70.3	80.0	79.0
HRNet-W32 [4]	HRNet-W32	384 × 288	28.5M	16.0G	74.9	92.5	82.8	71.3	80.9	80.1
HRNet-W48 [4]	HRNet-W48	384 × 288	63.6M	32.9G	75.5	92.5	83.3	71.9	81.5	80.5
HRFormer-B [29]	HRFormer	384 × 288	43.2M	26.8G	76.2	92.7	83.8	72.5	82.3	81.2
HRGCNet-W32(ours)	HRGCNet	384 × 288	29.6M	16.1G	77.9	93.6	84.8	74.8	82.9	80.6
HRGCNet-W48(ours)	HRGCNet	384 × 288	64.6M	32.9G	**78.3**	**93.6**	**85.7**	**75.3**	**83.5**	**81.2**

input size and obtains 76.6 AP parameters, which is superior to other methods with the same input size. (i) Compared with HRFormer-B, our computational complexity is high, but the gain is 1.8%. (ii) Compared with the TransPose-H-A6 and TokenPose-L/D24, AP has achieved 1.1% gain, but our GFLOPs are smaller than TransPose-H-A6. (iii) Compared with previous HRNet with the best performance, when the model sizes (#Params) and GFLOPs are similar, AP with input sizes of 256 × 192 for HRNe-w32 and HRNe-w48 increased by 2.2% and 2.3% respectively.

HRGCNet-w32 and HRGCNet-w48 with input size of 384 × 288 get AP scores of 77.8 and 78.4. Compared with HRNet-w32 and HRNet-w48 with input size of 384 × 288, our network has increased AP by 2.2% and 2.1%, respectively. At the same time, our network's number of parameters and GFLOPs have not increased much. Our network model increases the AP score by 1.2 compared to HRFormer. We believe that our HRGCNet can achieve better results by utilizing UDP or DARK schemes.

2) COCO test-dev: Table II shows a comparison of our pose estimation performance with existing methods. Our network model achieves 78.3 AP accuracy. Compared with HRNet, it does not increase much in model size or computational complexity but gains at least 1.3 gains. On the other hand, our small network model gains 1.4 compared to TransPose-H-A6, and our extensive network model gains 2.1 compared to HRFormer-B.

V. CONCLUSION

In this paper, we apply a combination of Bottleneck Blocks with the lightweight GC block to the high-resolution subnet of HRNet. It adds a global context to each location of the feature map, enriches the feature information in the high-resolution

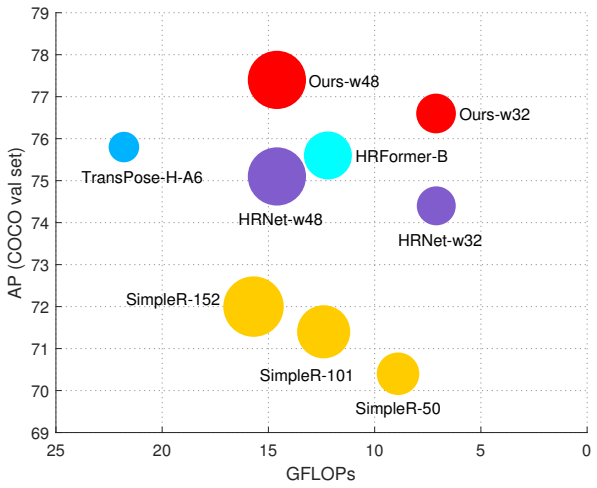


Fig. 3. The comparison of HRGCNet and SOTA methods on COCO val set. The input size of the model is 256 × 192, and the bubble size indicates the number of model parameters.

feature map, and improves model accuracy without increasing computational costs. Experiments in the COCO validation set show that the AP score is improved by 2.1 compared to HRNet when the number of parameters and computation are equal. Our method also yields good results on COCO train2017 datasets compared to those with the best performance.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 62172313.

References

[1] A. Cioppa, A. Deliege, S. Giancola, B. Ghanem, M. V. Droogenbroeck, R. Gade, and T. B. Moeslund, "A context-aware loss function for action spotting in soccer videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 126–13 136.

[2] F. Rea, A. Vignolo, A. Sciutti, and N. Noceti, "Human motion understanding for selecting action timing in collaborative human-robot interaction," *Frontiers in Robotics and AI*, vol. 6, p. 58, 2019.

[3] Y. Ding, D. Barath, J. Yang, H. Kong, and Z. Kukelova, "Globally optimal relative pose estimation with gravity prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 394–403.

[4] 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*, 2019, pp. 5693–5703.

[5] 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*, 2020, pp. 5386–5395.

[6] K. Li, S. Wang, X. Zhang, Y. Xu, W. Xu, and Z. Tu, "Pose recognition with cascade transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1944–1953.

[7] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "Vitpose: Simple vision transformer baselines for human pose estimation," *arXiv preprint arXiv:2204.12484*, 2022.

[8] D. Osokin, "Real-time 2d multi-person pose estimation on cpu: Lightweight openpose," *arXiv preprint arXiv:1811.12004*, 2018.

[9] 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*, 2021, pp. 10 440–10 450.

[10] Z. Zhang, J. Tang, and G. Wu, "Simple and lightweight human pose estimation," *arXiv preprint arXiv:1911.10346*, 2019.

[11] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 648–656.

[12] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler, "Joint training of a convolutional network and a graphical model for human pose estimation," *Advances in neural information processing systems*, vol. 27, 2014.

[13] 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*, 2016, pp. 770–778.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[15] 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)*, 2018, pp. 466–481.

[16] 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*, 2017, pp. 7291–7299.

[17] 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*, 2017, pp. 4903–4911.

[18] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[19] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3146–3154.

[20] Q. Wang, T. Wu, H. Zheng, and G. Guo, "Hierarchical pyramid diverse attention networks for face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8326–8335.

[21] W. Du, Y. Wang, and Y. Qiao, "Recurrent spatial-temporal attention network for action recognition in videos," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1347–1360, 2017.

[22] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1831–1840.

[23] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[25] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "Gcnet: Non-local networks meet squeeze-excitation networks and beyond," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[26] 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*, 2018, pp. 7103–7112.

[27] S. Yang, Z. Quan, M. Nie, and W. Yang, "Transpose: Keypoint localization via transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 802–11 812.

[28] Y. Li, S. Zhang, Z. Wang, S. Yang, W. Yang, S.-T. Xia, and E. Zhou, "Tokenpose: Learning keypoint tokens for human pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 313–11 322.

[29] Y. Yuan, R. Fu, L. Huang, W. Lin, C. Zhang, X. Chen, and J. Wang, "Hrformer: High-resolution transformer for dense prediction," *arXiv preprint arXiv:2110.09408*, 2021.

[30] S. Huang, M. Gong, and D. Tao, "A coarse-fine network for keypoint localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3028–3037.