Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.041538

ARTICLE

Full Scale-Aware Balanced High-Resolution Network for Multi-Person Pose Estimation

Shaohua Li, Haixiang Zhang*, Hanjie Ma, Jie Feng and Mingfeng Jiang

School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou, 310018, China

*Corresponding Author: Haixiang Zhang. Email: zhhx@zstu.edu.cn

Received: 26 April 2023 Accepted: 25 June 2023 Published: 08 October 2023

ABSTRACT

Scale variation is a major challenge in multi-person pose estimation. In scenes where persons are present at various distances, models tend to perform better on larger-scale persons, while the performance for smaller-scale persons often falls short of expectations. Therefore, effectively balancing the persons of different scales poses a significant challenge. So this paper proposes a new multi-person pose estimation model called FSA Net to improve the model's performance in complex scenes. Our model utilizes High-Resolution Network (HRNet) as the backbone and feeds the outputs of the last stage's four branches into the DCB module. The dilated convolution-based (DCB) module employs a parallel structure that incorporates dilated convolutions with different rates to expand the receptive field of each branch. Subsequently, the attention operation-based (AOB) module performs attention operations at both branch and channel levels to enhance high-frequency features and reduce the influence of noise. Finally, predictions are made using the heatmap representation. The model can recognize images with diverse scales and more complex semantic information. Experimental results demonstrate that FSA Net achieves competitive results on the MSCOCO and MPII datasets, validating the effectiveness of our proposed approach.

KEYWORDS

Computer vision; high-resolution network; human pose estimation

1 Introduction

Human pose estimation is a hot topic in computer vision, which involves simulating the human body's joints by predicting key points. In recent years, it has been extensively applied in various fields, including but not limited to autonomous driving [1–3], motion tracking [4–6], and intelligent surveillance [7–9].

In human body key point detection, two mainstream methods currently prevail: top-down and bottom-up. The former involves a multi-stage approach, typically detecting a single-person image using an object detector, followed by single-person pose estimation. The latter is a single-stage method primarily used for multi-person pose estimation, where the standard approach is to directly predict all key points in the image and subsequently perform classification and aggregation. The bottom-up method is more challenging, as larger-scale person features are typically more prominent and more accessible for the network to learn during training. In contrast, smaller-scale person features are inherently more ambiguous and susceptible to surrounding noise. As such, a key focus of research is developing methods to enable the network to perceive better targets of varying scales, which is also the primary focus of our study.

In multi-person pose estimation tasks, the design of the backbone network plays a crucial role due to the prevalent use of single-stage structures in mainstream multi-person pose estimation models. The depth, downsampling rate, and receptive field size of the network significantly impact its performance. Although HRNet [10] is a popular choice as the backbone, its larger network depth enhances its ability to handle large targets but potentially neglects small-scale targets. Moreover, an excessive number of downsampling layers can impair the network's capability to detect small-scale targets, thus affecting HRNet's performance and making it difficult to balance the detection of targets of varying scales.

In summary, we propose a bottom-up approach with HRNet as the backbone network, coupled with our meticulously designed post-processing modules, which we refer to as FSA Net. Fig. 1 illustrates the rough structure of FSA Net, which encompasses the DCB and AOB modules within the full-scale aware module (FSAM). In contrast to the previous practice of merging the four branches of the backbone network into one branch using a simple upsampling and addition method, our method addresses the semantic differences between them and considers the importance of each feature. Specifically, we input the outputs of the four branches into the DCB module, which comprises four parallel branches with dilated convolutions having different dilation rates of 9, 7, 5, and 3, respectively, to boost the receptive fields of various branches. After processing with the DCB module, the four feature maps of the branches are cross-stitched to form new feature maps of the four branches. Furthermore, attention operations are performed on the feature maps of each branch from both the branch and channel dimensions to enhance essential features and suppress redundant ones. Finally, we employ the heatmap representation to predict the final key points. The main contributions of this paper are summarized as follows:

- We present FSA Net, a high-resolution human pose estimation network that achieves scale-aware balance by paying attention to the performance of targets at different scales and extracting more complex semantic information during training. In contrast to other networks that tend to ignore the performance of small-scale targets, FSA Net can achieve a full-scale perception balance.
- We propose the DCB module, which covers targets at all scales through parallel structures and controls the receptive fields of different branches through dilated convolutions, thereby better perceiving targets at different scales.
- We propose the AOB module, which performs attention operations on feature maps of different branches after concatenating them. Unlike other models that simply add feature maps of different branches without considering their semantic differences, the AOB module can enhance the fusion ability of multi-scale features by strengthening important features and suppressing noise, thus improving the detection ability of small-scale targets.
- We have evaluated our proposed method on the widely used COCO Validation and test-dev datasets and attained remarkable results. Moreover, the feasibility of our approach was validated by conducting ablation experiments.

2 Related Studies

In single-person pose estimation, each person is detected separately, and there is no scale variation problem, making it much simpler than multi-person pose estimation. This chapter focuses on the development history of multi-person pose estimation. This paper first provides an overview of the overall development of multi-person pose estimation, followed by an exploration of the use of high-resolution networks in this task. Finally, the paper introduces attention mechanisms in multi-person pose estimation.

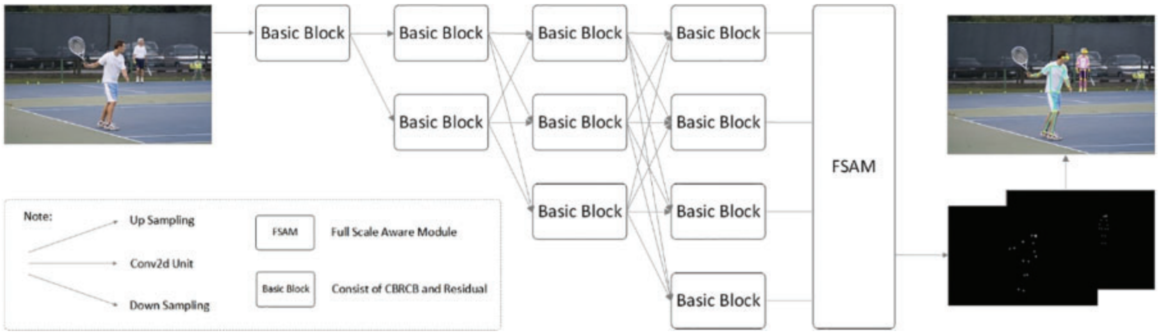

Figure 1: Simplified overall framework diagram of FSA Net

2.1 Mutil-Person Pose Estimation

Multi-person pose estimation involves predicting all the key points in an image and grouping them. Hourglass [11] proposed a cascading pyramid network that performs pose estimation by regressing key points directly. In Deepcut [12], the authors used Faster RCNN [13] for human detection, followed by the integer linear program method for pose estimation. Deepercut [14] employed ResNet [15] for body part extraction and Image Conditioned Pairwise Terms for key point prediction, improving accuracy and speed. Personlab [16] detected all the key points in an image using a box-free approach and then used greedy decoding to cluster the key points by combining the predicted offsets. OpenPose [17] used VGG [18] as the backbone network and proposed Part Affinity Fields to connect the key points. PifPaf [19] outperformed previous algorithms in low-resolution and heavily occluded scenes, and Pif predicts body parts, while Paf represents the relationships between them.

2.2 High-Resolution Network

Experimental results have shown that high-resolution feature maps are crucial for achieving superior detection performance in complex tasks such as human pose estimation and semantic segmentation. The proposal of HRNet has caused a huge response, as the authors employed a parallel connection of multiple sub-networks with varying resolutions, maintaining a high-resolution branch and performing repeated multi-scale fusion to enable each high-resolution feature map to receive information from other parallel branches of different resolutions repeatedly, thus generating rich high-resolution representations. Subsequently, the authors proposed HigherHRNet [20], which adopted a bottom-up approach using transpose convolution to expand the size of the last layer output feature map to half of the original image size, leading to improved performance. BalanceHRNet [21] introduced a balanced high-resolution module BHRM was proposed, along with a branch attention module for the branch fusion method to capture the importance of different branches. Multi-Stage HRNet [22] used a bottom-up approach to parallel multiple levels of HRNet and employed cross-stage feature fusion to further refine key point prediction. Wang et al. [23] enhanced the network's ability to capture global context information using switchable convolution operations. NHRNet [24] optimized HRNet by cutting high-level features in low-resolution branches and adding attention mechanisms in residual blocks. Khan et al. [25] proposed a network to address the challenges of multi-scale and global context feature loss and applied it to flood disaster response in high-resolution images.

2.3 Attention Mechanism

The attention mechanism has become a cornerstone in many computer vision tasks including image classification, object detection, instance segmentation, and pose estimation, owing to its remarkable simplicity and effectiveness. It enables networks to selectively focus on relevant information while ignoring irrelevant noise, thus resulting in improved performance. In recent years, several attention-based models have been proposed to enhance the feature fusion capability of the HRNet. For instance, SaMr-Net [26] introduced dilated convolutions and attention modules to promote multi-scale feature fusion. HR-ARNet [27] added a refinement attention module to the end of the HRNet for improved feature fusion. Improved HRNet [28] added a dual attention mechanism to parallel sub-networks to minimize interference caused by unrelated information. Meanwhile, Zhang et al. [29] enhanced the feature expression ability of each branch of HRNet by adding attention modules to different branches. X-HRNet [30] employed a one-dimensional spatial self-attention module to generate two one-dimensional heatmaps by projecting the two-dimensional heatmap onto the horizontal and vertical axes for keypoint prediction.

3 Proposed Method

In this section, we present a comprehensive exposition of FSA Net. Fig. 1 provides a schematic representation of the network architecture, while Fig. 2 offers a more intuitive glimpse into the method's implementation process. Specifically, we employ HRNet as the backbone network, assuming the input image is $X \in \mathbb{R}^{3\times H\times W}$, after the feature extraction by HRNet, the four branch features are obtained: $X_{b1} \in \mathbb{R}^{C_1\times H_1\times W_1}$, $X_{b2} \in \mathbb{R}^{C_2\times H_2\times W_2}$, $X_{b3} \in \mathbb{R}^{C_3\times H_3\times W_3}$ and $X_{b4} \in \mathbb{R}^{C_4\times H_4\times W_4}$. The size of the feature maps in branch b_1 is $\frac{1}{4}$ of the original image size, while the feature maps sizes in branches b_2 to b_4 are $\frac{1}{8}$, $\frac{1}{16}$ and $\frac{1}{32}$ of the original image size, respectively. The source paper uses the highest-resolution branch as the network's output, followed by post-processing. In this paper, we utilize the outputs from all four branches.

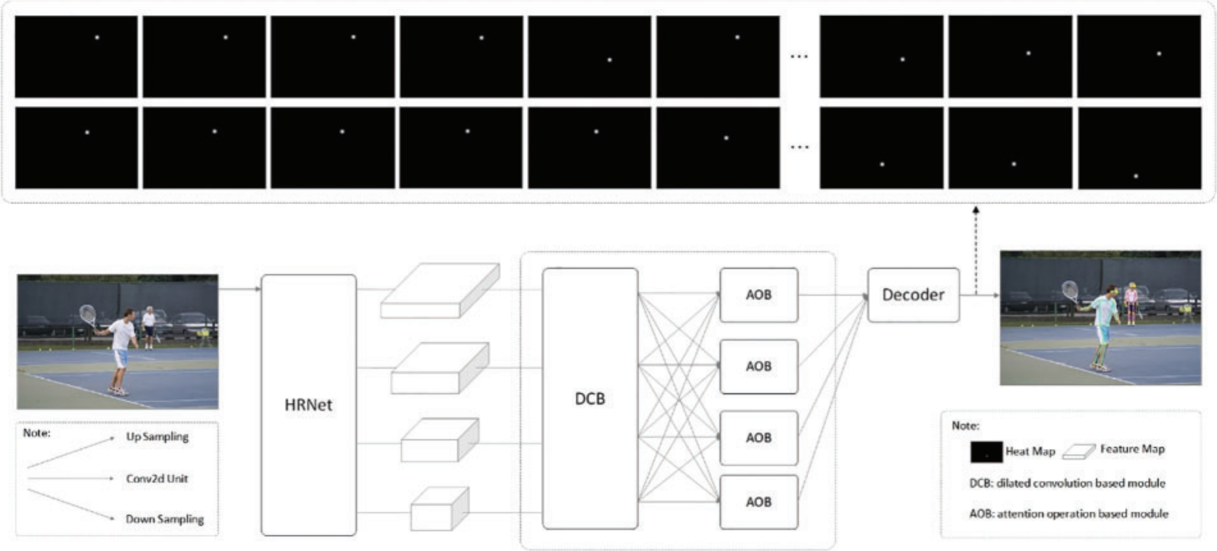

Figure 2: Overall framework diagram of SSA Net

3.1 Dilated Convolution Based Module

Multi-person images are fed into our network, and after feature extraction by the backbone network, the feature maps of four branches are outputted. These feature maps are then individually passed through the DCB module, consisting of four parallel branches with different dilation rates of dilated convolution, as illustrated in Fig. 3. Specifically, each branch's output feature maps undergo bottleneck processing similar to that of ResNet [15], consisting of 1 × 1 convolution, dilated convolution with different dilation rates, and 1 × 1 convolution again. The final bottleneck uses a residual operation instead of an activation function.

In HRNet, limited receptive field and excessive downsampling severely affect the detection ability of small objects. Therefore, we propose the DCB module, which accurately controls the receptive field of each branch through a parallel structure. For the branch with the highest resolution, we use dilated convolution with a dilation rate of 9 to expand the receptive field of each pixel. For the branch with the lowest resolution, we limit its receptive field to a smaller range to better perceive small-scale targets, enhancing the perception ability of targets of various scales.

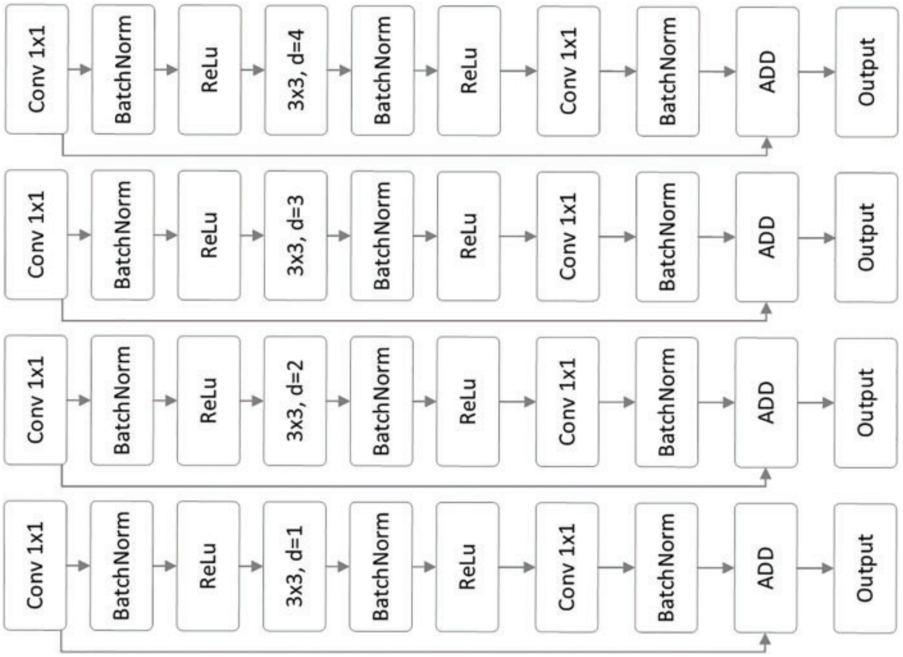

Figure 3: Schematic diagram of the DCB module

3.2 Attention Operation-Based Module

After undergoing processing by the DCB module, the network can enhance the perception of objects at all scales. However, due to their relatively blurred features, small-scale persons are easily affected by noise. To address this issue, we introduce the AOB module, which performs attention operations on each branch and channel to emphasize important features and suppress noise. This results in significant improvements in detection performance, particularly for small-scale objects. Attention operations on each branch also highlight the semantic differences between branches, enabling the network to better attend to targets of different scales.

Specifically, as shown in Fig. 4. The AOB module takes the output feature maps of the DCB module, denoted as w_1, w_2, w_3, w_4, as input. These feature maps are first subjected to cross-splicing to generate $branchw_1, branchw_2, branchw_3, branchw_4$, as shown in Eq. (1):

$$branch\ w_i = \text{Concatenate} \sum\nolimits_{j=1}^{4} w_{(i,j)} \tag{1}$$

Where i denotes the i_{th} branch, and j denotes the number of used branches. Then the attention operation is performed on $branch\ w_i$ from branch and channel dimensions. Firstly, each branch is processed by adaptive average pooling and adaptive max pooling, and then sum the results to obtain the attention score, after which the obtained attention score is applied to $branch\ w_i$ to readjust the weights, and the calculation process is shown in Eq. (2):

$$feature\ w_i = branch\ w_i \times [AAP\,(branch\ w_i) + AMP\,(branch\ w_i)] \tag{2}$$

Where AAP stands for adaptive average pooling operation, and AMP stands for adaptive max pooling operation. Then feature fusion is performed on the feature maps after the attention operation to regenerate the four branchs feature maps, and the calculation process is shown in Eq. (3):

$$w_i' = \sum\nolimits_{j=1}^{4} feature\ w_{(i,j)} \tag{3}$$

where i denotes the i_{th} branch, and j denotes the number of branches to be used.

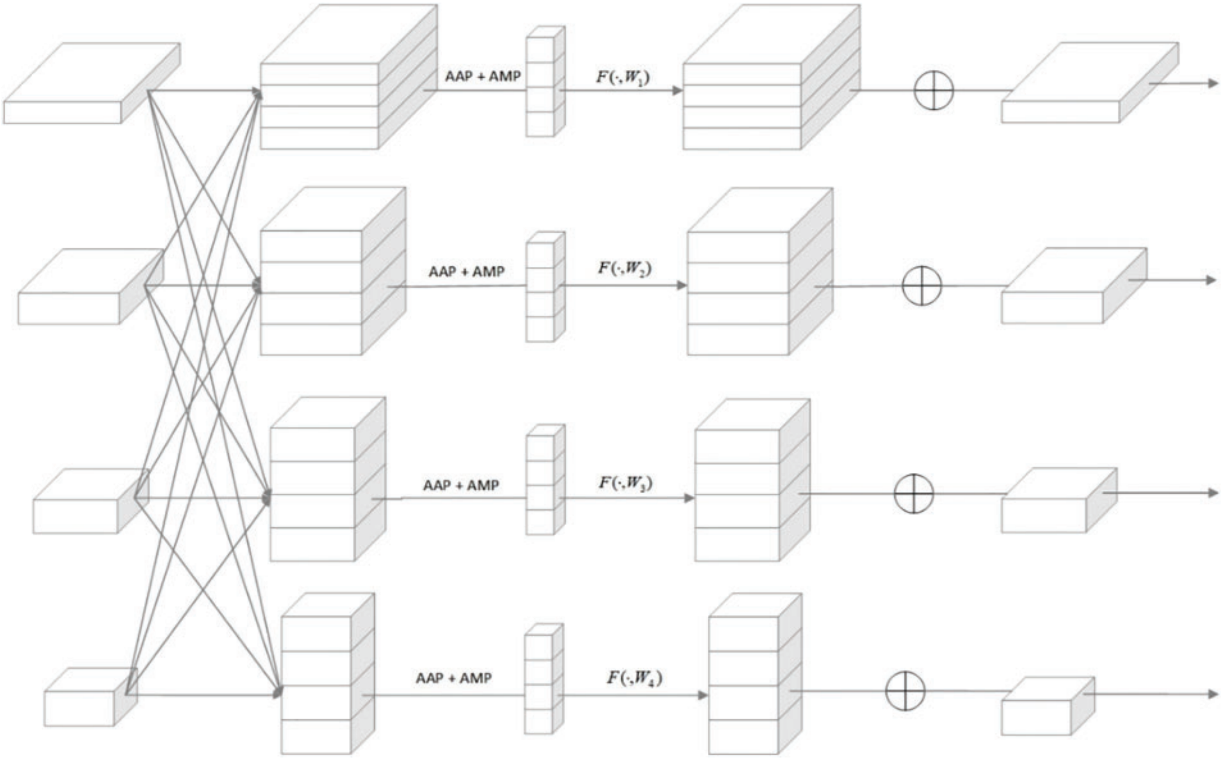

Figure 4: Schematic diagram of the AOB module

3.3 Decoder Module

In the last stage of HRNet, only the branch with the highest resolution is selected as the final output, while the contributions of other branches are ignored. This approach may hurt the performance of targets at other scales. To address this issue, in our decoding stage, we fuse the outputs of all branches by upsampling, generating a heatmap the same size as the original image and with a 1/4 resolution. This heatmap integrates features of all resolutions and can perceive the key points in different scale persons. The predicted coordinates are then decoded, and the L2 loss function is utilized to calculate the loss by comparing the estimated heatmap with the ground truth heatmap. During the loss calculation in the training process, since we have (the number of key points) channels, we compute the loss by calculating the L2 norm of the corresponding key points' indices in each channel. As the labels are generated as heatmaps representing a small region, the loss calculation measures the distance (L2 distance) between the predicted key points and the key points in the labeled region.

4 Experiments

4.1 Experimental Details

In the FSA Net, we adopted a bottom-up approach for multi-person pose estimation. We set the initial learning rate to 0.001 during the experiments and used the Adam optimizer to obtain gradients. The total number of epochs was set to 300, and we trained and evaluated the network on the widely used COCO dataset. Due to the use of heatmap-based methods, a significant amount of memory was required. The experimental hardware and software environment is presented in Table 1.

Table 1: The hardware and software configuration environment for the experiments

Hardware	CPU	Intel(R) Xeon(R) CPU E5-2678 v3 @2.50GHz $*$ 48
	GPU	NVIDIA GeForce RTX 3090 24G $*$ 8
Software	OS	Linux Ubuntu 20.04.5 LTS
	Python version	Python 3.7.0
	Pytorch version	Pytorch 1.13.1
	Cuda version	Cuda11.6 + Cudnn8.3.2

4.2 Datasets and Evaluation Metrics

The dataset employed in this experiment is the widely used MSCOCO dataset in human pose estimation. MSCOCO is a large-scale, multi-purpose dataset developed by Microsoft that plays an essential role in mainstream visual tasks. It contains over 200 k images, with 250 k annotated instances of human key points. The MSCOCO training set consists of 118 k images. In contrast, the test set comprises two subsets: COCO Validation, including 5 k images mainly used for simple testing and ablation experiments, and COCO test-dev, including 20 k images mainly used for online testing and fair comparison with state-of-the-art models. The primary evaluation metrics in the COCO dataset are average precision (AP) and average recall (AR), which are calculated according to the following formulas:

$$AP_t = \frac{\sum_p \delta(OKS > t)}{\sum_p 1} \tag{4}$$

where p is the number of detected human instances, and t is the threshold to refine the evaluation index.

When t is taken as 0.5 and 0.75, it is noted as AP^{50} and AP^{75}, and OKS is object key point similarity, which is calculated as follows:

$$\text{OKS} = \frac{\sum_i exp^{\left(-\frac{d_i^2}{2s^2k_i^2}\right)}\delta(V_i > 0)}{\sum_i \delta(V_i > 0)} \tag{5}$$

where i denotes the i_{th} key point, d_i denotes the Euclidean distance between the true value and the predicted value, s is the area of human instance, V_i is whether the recognition instance is visible or not, and δ is the regularization parameter of the key point. When $32^2 < s^2 < 96^2$ and $96^2 < s^2$, we write AP^M and AP^L, and similarly AR can be written as AR^{50}, AR^{75}, AR^M, AR^L.

In addition, we also constructed the Tiny Validation dataset, which is a subset of the COCO Validation dataset. In this dataset, we labeled images that contain persons with an area smaller than 80^2 as "images with small individuals." After filtering through the 5000 images, we found that only 361 images met this criterion. We curated these images to form a new dataset. The purpose of using this dataset is to evaluate the network's detection performance specifically for small individuals.

4.3 Quantitative Experimental Results

In the study of full-scale aware balance, detecting small-scale person poses is the most significant challenge. Traditional pose estimation networks tend to focus more on medium to large-scale persons and may overlook the performance of small-scale persons. Therefore, we first evaluated FSANet's detection capability for small persons on the Tiny Validation dataset. Table 2 shows that HigherHRNet [20] exhibits significant performance degradation on the Tiny Validation dataset, with an average loss of 20% in AP. This confirms the inference mentioned earlier. Additionally, it can be seen that FSA Net shows a substantial performance improvement, and FSANet-W48 achieves an AP of 60.3%, surpassing HigherHRNet-W48 and other bottom-up models. This demonstrates the significant enhancement of our proposed FSA Net in detecting small-scale persons.

Table 2: Comparison with mainstream bottom-up models on the tiny validation dataset

Method	Backbone	InputSize	AP
HigherHRNet-W32 [20]	HRNetW32	512	47.4($\downarrow$19.7)
HigherHRNet-W48 [20]	HRNetW48	640	49.0($\downarrow$20.9)
FSANet-W32	HRNetW32	512	59.1($\downarrow$9.5)
FSANet-W48	HRNetW48	640	60.3($\downarrow$10.9)

In Table 3, we present the results of testing FSA Net on the widely used COCO Validation dataset, where we achieved significant improvements across multiple metrics. Specifically, when utilizing HRNet-W32 as the backbone network, FSA Net outperformed HigherHRNet by 1.5% and achieved even greater improvement compared to networks such as the method [15]. Furthermore, when the backbone network was HRNet-W48, the network's AP reached 71.2, which was 1.3% higher than HigherHRNet. Notably, we observed varying degrees of improvement in both AP^M and AP^L, underscoring the enhanced performance of our method across different scales.

Table 3: Comparison with mainstream bottom-up models on the COCO Validation dataset, bold is the best result in each column, HG is the hourglass, DLA is deep layer aggregation, W32 is HRNet-W32 and W48 is HRNet-W48

Method	Input size	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
CenterNet-DLA [31]	512	58.9	–	–	–	–
CenterNet-HG [31]	512	64.0	–	–	–	–
PifPaf [18]	–	67.4	–	–	–	–
Personlab [15]	601	54.1	76.4	57.7	40.6	73.3
Personlab [15]	1401	66.5	86.2	71.9	62.3	73.2
HigherHRNet-W32 [20]	512	67.1	86.2	73.0		
FSANet-W32	512	68.6	87.1	74.2	64.2	76.8
HigherHRNet-W48 [20]	640	69.9	87.2	76.1		
FSANet-W48	640	**71.2**	**88.5**	**77.4**	**66.0**	**79.1**

In Table 4, we further evaluated our model on the COCO test-dev dataset to strengthen its persuasiveness. Our experiments show that when HRNetW48 was used as the backbone network, our method outperformed other mainstream bottom-up models, achieving an AP of 70.8. It is worth noting that HigherHRNet and our work share similar ideas of enlarging the feature map area to better perceive small-scale objects by upsampling the last output of HRNet through transpose convolution. However, FSA Net focuses on the performance of each branch, expands the receptive field through parallel-structured dilated convolution, and eliminates noise through branch attention operations. Our experimental results demonstrate the clear superiority of our proposed approach. In addition, since we process all four branches, the parameter count and computational complexity (GFLOPs) of FSANet are slightly higher than HrHRNet but significantly better than other models such as Hourglass.

Table 4: Comparison with mainstream bottom-up models on COCO test-dev dataset, bold is the best result in each column, BU is the bottom-up method, HrHRNet is HigherHRNet, HG is Hourglass, R-152 is ResNet152, H-W32 is HRNetW32

Method	Backbone	Input size	#Param	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
Openpose [16]	–	–	–	–	61.8	84.9	67.5	57.1	68.2
Hourglass [10]	HG	512	277.8	206.9	56.6	81.8	61.8	49.8	67.0
Personlab [15]	R-152	1401	68.7	405.5	66.5	88.0	72.6	62.4	72.3
Pifpaf [18]	–	–	–	–	66.7	–	–	62.4	72.9
SPM [32]	–	–	–	–	66.9	88.5	72.9	62.6	73.1
BU HRNet [19]	H-W32	512	–	–	64.1	86.3	70.4	57.4	73.9
HrHRNet [20]	H-W32	512	28.5	47.9	66.4	87.5	72.8	61.2	74.2
FSANet	H-W32	512	31.7	49.1	67.9	88.3	74.7	62.4	75.3
HrHRNet [20]	H-W48	640	63.8	154.3	68.4	88.2	75.1	64.4	74.2
FSANet	H-W48	640	67.2	158.5	**70.8**	**89.1**	**77.6**	**65.9**	**77.4**

4.4 Qualitative Experimental Results

To visually demonstrate the effectiveness of our method, we have provided experimental results in Figs. 5 and 6. Specifically, we conducted tests on the COCO Validation dataset with HRNetW48 as the backbone network. As shown in Fig. 5. Our model can easily handle scenarios with rich scale information. Moreover, as illustrated in Fig. 6. The left is the original image, the middle is HigherHRNet, and the right is FSA Net. It can be seen that in complex scenes, such as self-occlusion of people, foreground occlusion, the size of targets is too small, and complex semantic information, our model predicts more accurately and reasonably than other models.

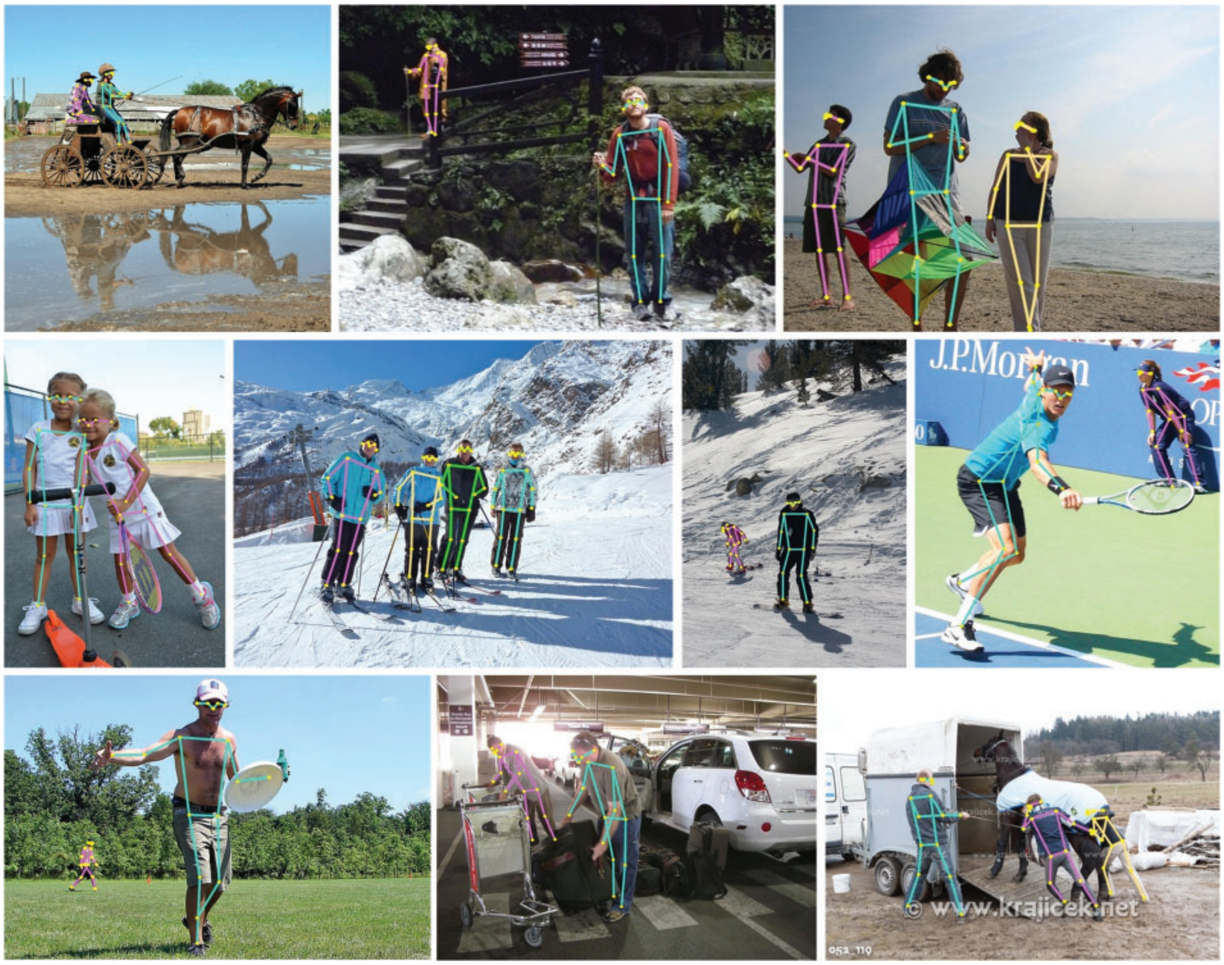

Figure 5: Visualization of the results of FSA Net on the COCO Validation dataset

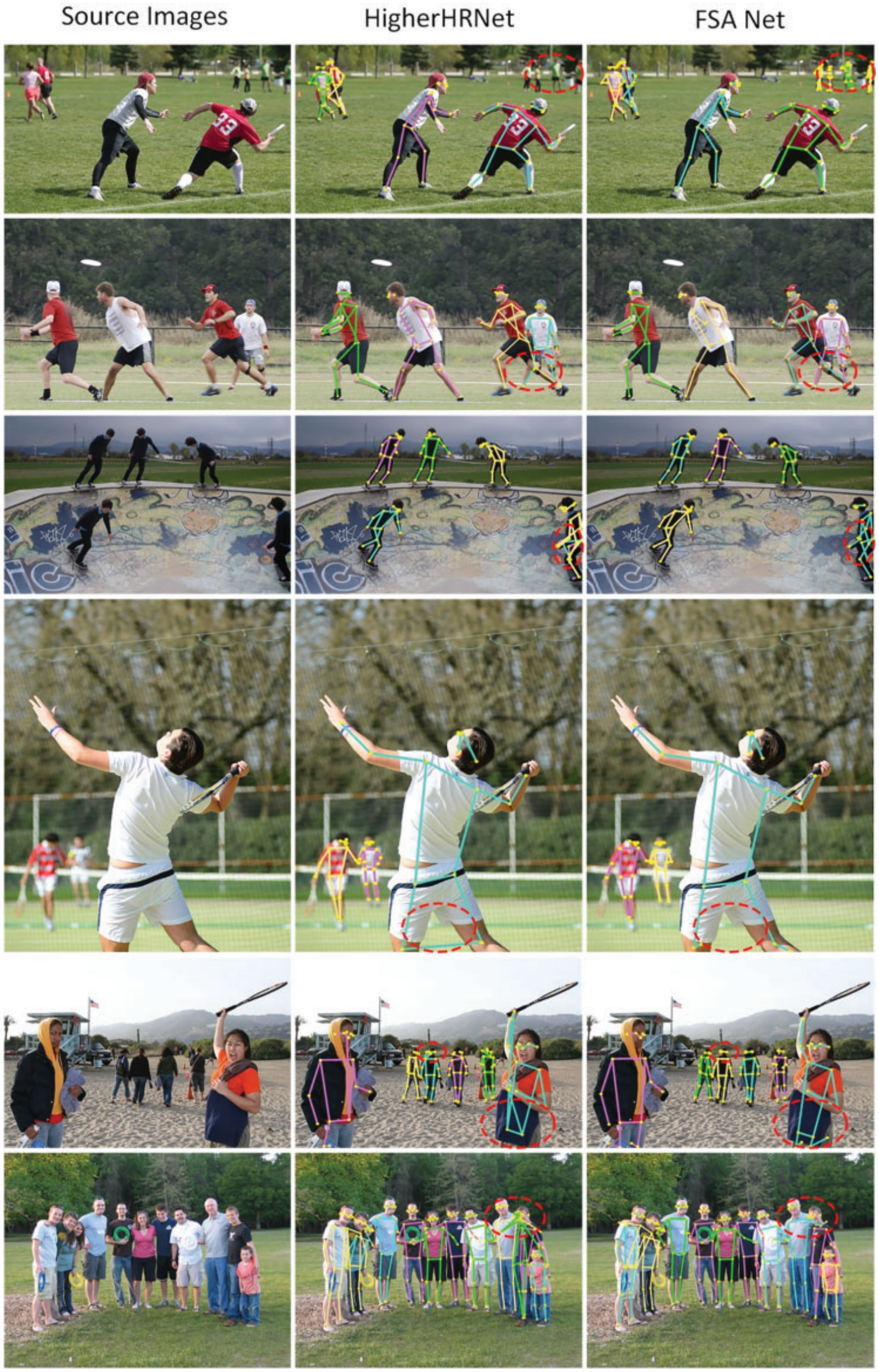

Figure 6: Qualitative comparison of FSA Net with other state-of-the-art bottom-up models on the COCO Validation dataset

4.5 Ablation Experiments

Our method builds upon the model HRNet by incorporating the DCB and AOB modules. To evaluate the impact of each module, we performed extensive ablation experiments on the COCO Validation dataset in Table 5, adopting the identical training strategy as in previous work. The results reveal that the DCB module boosts AP by 2.5% compared to the baseline model, while the AOB module achieves a significant increase of 3.1% in AP. Both modules have contributed to the overall performance improvement, with the AOB module exhibiting a greater contribution. When the two modules are jointly applied, the overall AP is further boosted by 4.2%, attesting to the remarkable effectiveness of our proposed approach.

Table 5: Ablation experiments on COCO validation, Ba is Baseline

Method	HRNet	DCB	AOB	AP
HRNet	✓			64.4
HRNet + DCB	✓	✓		66.9(↑2.5%)
HRNet + AOB	✓		✓	67.5(↑3.1%)
FSANet	✓	✓	✓	68.6(↑4.2%)

5 Conclusion

This study introduces a novel multi-person high-resolution pose estimation network with full-scale perception balance. Our approach improves upon the HRNet baseline by incorporating DCB and AOB modules. The DCB module expands the receptive field of each branch's results, while the AOB module performs attention operations to enhance feature representation ability. We evaluate our method on popular datasets and outperform the baseline model and similar networks. Additionally, we visualize the experimental results to provide a more intuitive understanding of the performance improvements of our FSA Net. In our initial attempts, we drew inspiration from [33] and used dilation rates of 4, 3, 2, and 1 for the dilated convolutions in the DCB module. However, the experimental results were not satisfactory. Upon analysis, we realized that precise localization of key points is crucial in pose estimation. Therefore, we needed to use larger dilation rates to achieve a larger receptive field. Ultimately, we achieved success with dilation rates of 9, 7, 5, and 3. Although FSA Net has achieved impressive results, it still faces some challenges. As shown in Table 2, FSA Net can be considered to have achieved a scale-awareness balance compared to other bottom-up networks, but it has not fully achieved a scale-aware balance. There is still significant room for improvement in the performance on small-scale targets, which will be the focus of our future research.

Acknowledgement: Thanks for the full guidance and supervision from Associate Professor Haixiang Zhang, and the financial support from Professor Mingfeng Jiang, and we would like to express our gratitude to the esteemed editors and reviewers for their invaluable contributions to this article.

Funding Statement: This work is supported in part by the National Natural Science Foundation of China 61672466 62011530130, Joint Fund of Zhejiang Provincial Natural Science Foundation LSZ19F010001.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Shaohua Li; data collection: Shaohua Li; analysis and interpretation of results: Haixiang

Zhang, Hanjie Ma, Jie Feng; draft manuscript preparation: Mingfeng Jiang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: This experiment uses the public COCO dataset, which is directly available, and since there are subsequent algorithm improvements, the code can be obtained from the corresponding author under reasonable circumstances.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present research.

References

[1] R. Gu, G. Wang and J. N. Hwang, "Efficient multi-person hierarchical 3D pose estimation for autonomous driving," in *IEEE Conf. on Multimedia Information Processing and Retrieval*, San Jose, CA, USA, pp. 163–168, 2019.

[2] J. Zheng, X. Shi, A. Gorban, J. Mao, Y. Song *et al.,* "Multi-modal 3D human pose estimation with 2D weak supervision in autonomous driving," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, New Orleans, LA, USA, pp. 4478–4487, 2022.

[3] A. Dhall, D. Dai and L. Van Gool, "Real-time 3D traffic cone detection for autonomous driving," in *IEEE Intelligent Vehicles Symp.*, Paris, France, pp. 494–501, 2019.

[4] F. Cordella, F. Di Corato, G. Loianno, B. Siciliano and L. Zollo, "Robust pose estimation algorithm for wrist motion tracking," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, Tokyo, Japan, pp. 3746–3751, 2013.

[5] S. Sridhar, A. Oulasvirta and C. Theobalt, "Interactive markerless articulated hand motion tracking using RGB and depth data," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Sydney, Australia, pp. 2456–2463, 2013.

[6] A. Kyme, S. Se, S. Meikle, G. Angelis, W. Ryder *et al.,* "Markerless motion tracking of awake animals in positron emission tomography," *IEEE Transactions on Medical Imaging*, vol. 33, no. 11, pp. 2180–2190, 2014.

[7] W. Rahmaniar, Q. M.ul Haq and T. L. Lin, "Wide range head pose estimation using a single RGB camera for intelligent surveillance," *IEEE Sensors Journal*, vol. 22, no. 11, pp. 11112–11121, 2022.

[8] D. Li, X. Chen, Z. Zhang and K. Huang, "Pose guided deep model for pedestrian attribute recognition in surveillance scenarios," in *IEEE Int. Conf. on Multimedia and Expo*, San Diego, CA, USA, pp. 1–6, 2018.

[9] M. Cormier, A. Clepe, A. Specker and J. Beyerer, "Where are we with human pose estimation in real-world surveillance?," in *Proc. of the IEEE/CVF Winter Conf. on Applications of Computer Vision*, Waikoloa, HI, USA, pp. 591–601, 2022.

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

[11] A. Newell, K. Yang and J. Deng, "Stacked hourglass networks for human pose estimation," in *Computer Vision–ECCV 2016: 14th European Conf.*, Amsterdam, Netherlands, pp. 483–499, 2016.

[12] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka *et al.,* "Deepcut: Joint subset partition and labeling for multi person pose estimation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 4929–4937, 2016.

[13] R. Girshick, "Fast R-CNN," in *Proc. of the IEEE Int. Conf. on Computer Vision*, Santiago, Chile, pp. 1440–1448, 2015.

[14] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," in *Computer Vision-ECCV 2016: 14th European Conf.*, Amsterdam, Netherlands, pp. 34–50, 2016.

[15] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, USA, pp. 770–778, 2016.

[16] G. Papandreou, T. Zhu, L. C. Chen, S. Gidaris, J. Tompson *et al.,* "Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model," in *Proc. of the European Conf. on Computer Vision*, Munich, Germany, pp. 269–286, 2018.

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

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

[19] S. Kreiss, L. Bertoni and A. Alahi, "Pifpaf: Composite fields for human pose estimation," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Long Beach, USA, pp. 11977–11986, 2019.

[20] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang *et al.,* "HigherhrNet: Scale-aware representation learning for bottom-up human pose estimation," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Seattle, USA, pp. 5386–5395, 2020.

[21] Y. Li, S. Jia and Q. Li, "BalanceHRNet: An effective network for bottom-up human pose estimation," *Neural Networks*, vol. 161, no. 893–6080, pp. 297–305, 2023.

[22] J. Huang, Z. Zhu and G. Huang, "Multi-stage HRNet: Multiple stage high-resolution network for human pose estimation," arXiv:1910.05901, 2019.

[23] R. Wang, C. Huang and X. Wang, "Global relation reasoning graph convolutional networks for human pose estimation," *IEEE Access*, vol. 8, pp. 38472–38480, 2020.

[24] S. Yang, D. He, Q. Li, Z. Li, J. Wang *et al.,* "Human pose estimation based on SNHRNet," in *Annual Int. Conf. on Network and Information Systems for Computers*, Shanghai, China, pp. 576–582, 2022.

[25] S. D. Khan and S. Basalamah, "Multi-scale and context-aware framework for flood segmentation in post-disaster high resolution aerial images," *Remote Sensing*, vol. 15, no. 2208, pp. 2072–4292, 2023.

[26] H. Yang, L. Guo, X. Wu and Y. Zhang, "Scale-aware attention-based multi-resolution representation for multi-person pose estimation," *Multimedia Systems*, vol. 28, no. 1, pp. 57–67, 2022.

[27] X. Wang, J. Tong and R. Wang, "Attention refined network for human pose estimation," *Neural Processing Letters*, vol. 53, no. 4, pp. 2853–2872, 2021.

[28] Y. Bao, M. Zhang and X. Guo, "Human pose estimation based on improved high resolution network," *Journal of Physics: Conference Series*, vol. 1961, no. 1, pp. 12060, 2021.

[29] C. Zhang, N. He, Q. Sun, X. Yin and K. Lu, "Human pose estimation based on attention multi-resolution network," in *Proc. of the 2021 Int. Conf. on Multimedia Retrieval*, Taipei, pp. 682–687, 2021.

[30] 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 Int. Conf. on Multimedia and Expo*, Taipei, pp. 1–6, 2022.

[31] X. Zhou, D. Wang and P. Krähenbühl, "Objects as points," arXiv:1904.07850, 2019.

[32] X. Nie, J. Feng, J. Zhang and S. Yan, "Single-stage multi-person pose machines," in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*, Long Beach, USA, pp. 6951–6960, 2019.

[33] Y. Li, Y. Chen, N. Wang and Z. Zhang, "Scale-aware trident networks for object detection," in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*, Seoul, Korea (South), pp. 6054–6063, 2019.