High-Resolution Network with Attention Module for Human Pose Estimation

Tien-Dat Tran, Xuan-Thuy Vo, Duy-Linh Nguyen and Kang-Hyun Jo
School of Electrical Engineering, University of Ulsan
Ulsan (44610), South Korea
Email: (tdat,xthuy)@islab.ulsan.ac.kr, ndlinh301@mail.ulsan.ac.kr, acejo@ulsan.ac.kr

***Abstract*—Convolution neural networks (CNNs) have achieved the highest performance today not only for human posture prediction but also for other machine vision tasks (e.g., object identification, semantic segmentation, images classification). Furthermore, the Attention Module demonstrates their superiority over other conventional networks (AM). As a result, this work focuses on a useful feed-forward AM for CNNs. First, following a stage in the backbone network, feed the feature map into the attention module, which is separated into two dimensions: channel and spatial. The AM then multiplies these two feature maps and passes them on to the next level in the backbone. The network can collect more information in long-distance dependencies (channels) and geographical data, resulting in higher precision efficiency. Our experimental results would also show a difference between the employment of the attention module and current methodologies. As a result of the switch to a High-resolution network (HRNet), the predicted joint heatmap keeps accuracy while reducing the number of parameters compared to the baseline-CNN backbone. In terms of AP, the suggested design outperforms the baseline-HRNet by 2.0 points. Furthermore, the proposed network was trained using the COCO 2017 benchmarks, which are currently available as an open dataset.**

***Index Terms*—machine learning, high-resolution network, attention module, human pose estimation.**

I. Introduction

In contemporary world nowadays, 2D human pose estimation plays an important but challenging function in computer vision, serving numerous objectives such as human re-identification [1], [2], activity recognition [3], [4], human pose estimation [5], [6] or 3D human pose estimation [7], [8]. Human pose's main goal is to recognize bodily sections for human body keypoint. Channel and spatial background are vital in improving the precision of key point regression. As a result, this research will concentrate on how to teach the network get better attention information.

Deep convolutional of neural networks have recently attained out standing performance, according to recent breakthroughs. Most existing approaches route the input through a network, which is generally made up of high-to-low resolution subnetworks connected in series, before increasing the resolution. Hourglass [9], for example, restores high resolution using a symmetric low-to-high process. SimpleBaseline [10] generates high-resolution representations using a few transposed convolution layers. Furthermore, dilated convolutions are employed to enlarge the latter layers of a high-to-low resolution network (e.g., VGGNet or ResNet) [11], [12].

Deep convolution of neural networks has now stored significant advances in human posture [13], [14]. These networks, however, still have a lot of challenges to sort out. First and foremost, how can accuracy be improved in various types of networks? (e.g., real-time network, accuracy network). Second, while updating or modifying a network, it is frequently necessary to examine its speed. Last but not least, the present network must improve accuracy while remaining as quick as feasible. This research describes an unique network and the attention module's dependability in terms of speed and accuracy. The suggested experiment compares using and not using the attention module. The experiment also differs from the Simple Baseline [10] experiment, which inactive the attention mechanism and for upsampling, it instead used the transpose convolution [15]. The proposed method would focus on how productive and economical each network situation is.

In particular, proposed technique was established a simple fine-tune attention module [16], which demonstrated a considerable improvement in mean Average Precision (mAP). Inspired by VGG16 [11], the suggested network attempts to enhance the spatial attention module (SAM) by employing two 3×3 convolution layers rather than a 7×7 convolution layer. The network maintains the mAP while lowering the implementation cost by using 3×3 kernel. Furthermore, the number of parameters was reduced, resulting in an increase in network speed. To understand clearly about AM, proposed network increase 4.7 point in Average Precision for precision and only increase around 16.5 percent of number parameters, which contrasted with the Attention mechanism standard [16] when utilized High-Resolution Network [17] as a backbone network. This research offers a novel network attention module that can readily react to a variety of difficulties in numerous applications, such as object identification, images classification, and human position estimation. The proposed method computes joint human pose estimations based on feature map recovery using an up-sampling method.

II. Related work

2D-Human Pose Estimation: The most important aspect of human pose estimation is key-point detection and its interaction with geographical data, Deeppose [18], Simple baseline makes use of joint prediction using an end-to-end architecture with a larger restriction. Later, Newell with the Stacked hourglass network [9] reduces the amount of settings

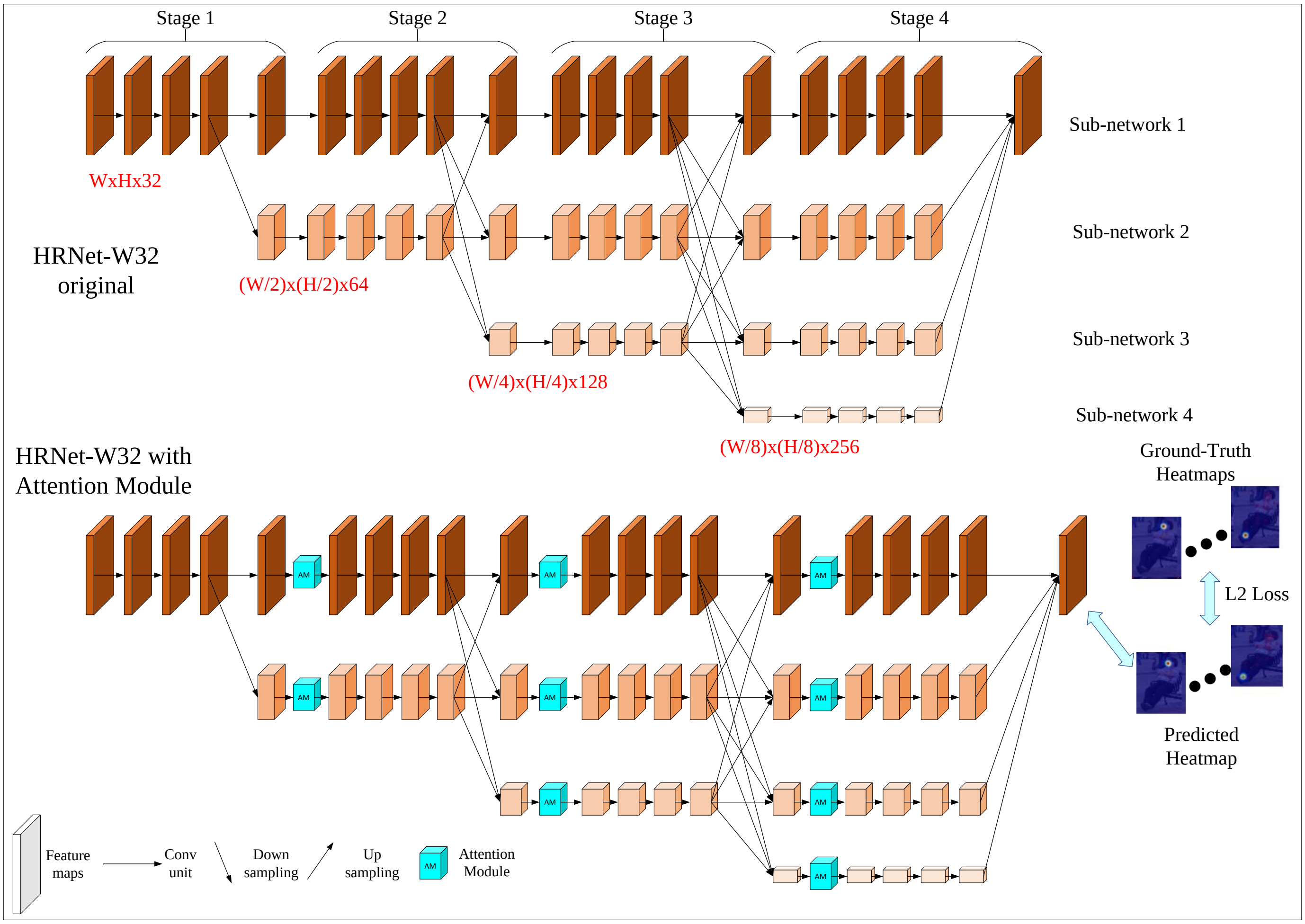

Fig. 1. Demonstrate the outline of the proposed 2D-human-pose estimation architecture. The proposed approach divided the network into 4 stages, each stage was connected by an attention module.

while maintaining great accuracy. Nowadays, Sun with the High-Resolution network [17] maintains the high-resolution map from beginning to end to keep the high-level feature for the network until the end. To represent local joints, all of the approaches employed Gaussian distributions. After that, a convolution neural architecture was utilized to predict human posture estimation. To minimize employment costs, they must shrink the quantity of parameters , and using appropriate attention approaches will reduce the network's parameter. As a result, the suggested strategy focuses on the employed attention module while increasing accuracy and decreasing the number of parameters.

On the other side, for increasing network performance, a 3×3 kernel size outperforms a 7×7 kernel size. However, in certain more sophisticated and expensive architectures, the 7×7 kernel size provides more precision. In comparison, our attention module gives a sufficient perspective for network design, with a limited number of parameters and high speed or a larger number of parameters and lower speed. The article then demonstrates how the attention module will function in each procedure and outcome.

High resolution network: Most convolutional neural networks for keypoint heatmap estimation are composed of a stem subnetwork, similar to a classification network, that decreases the resolution, a main body that produces representations with the same resolution as its input, and a regressor that estimates the heatmaps where the joint positions are estimated and then transformed in original resolution. Keeping the full resolution give the network get better accuracy. The main body primarily employs a high-to-low and low-to-high structure, which may be supplemented by multi-scale fusion and intermediate (deep) supervision.

In parallel, High Resolution architecture connects high-to-low subnetworks. It keeps high-resolution representations throughout the process, allowing for spatially exact heatmap estimate. It produces consistent high-resolution representations by repeatedly merging the representations created by the high-to-low subnetworks. Our technique differs from most previous efforts in that it requires a distinct low-to-high upsampling procedure as well as aggregate low-level and high-level feature map. Without the need of intermediate heatmaps supervision, the technique is superior in joint identification accuracy and efficient in computing complexity and parameters.

Attention mechanism: Human visualization is vital in computer vision, and a variety of focus processing methods are being made to improve the efficiency of CNNs. Wang et al. [19] also proposed a non-local network for gathering long-distance interdependence. SKNet [20] integrated the SENet Channel Focus Module with the Inception Multi-Branch Convolution, which was inspired by SENet [21] and Inception [22]. Furthermore, the Module for geographical attention is derived from Google's STN [23], which gathers the background data of the feature maps. Furthermore, the attention module provides several benefits for saliency detection, multi-label categorization, and individual identification.

The suggested approach in this research was motivated by the CBAM architecture [24] to create the productive in the middle of both spatial and channel module by employing element-wise multiplication. Following that, the tensors adds to the previous tensors to merge the old and latest data from the Attention block.

III. Methodology

A. Network architecture

Backbone network: Our architecture used a backbone that includes HRNet-W32 and HRNet-W48 [17], as shown in Figure 1 for a full architecture. Each HRNet has four phases, which include residual blocks and connections. The original input RGB image shrinks the size to 256×192 (HRNet-W32, HRNet-W48), the tensor traverse each pillar layer, and the initial resolution of $H \times W$ drops two times for every stage. Finally, after travelling down the backbone, the function map's dimension is decreased to $\frac{W}{16} \times \frac{H}{16}$ with 256 channels at the last bottom layer of network. However, the backbone network will only use the first subnetwork which keeps the size is $W \times H$ until the end of regression. Furthermore, the dimension of the channels got doubled at each stage. It progresses from 32 after the beginning stage to 256 in the final stage. The baseline network's job is to accumulate useful data from extract feature maps and transmit them to the Training System, which uses cross entropy loss to predict human joints.

After extracting the helpful data from the backbone architecture, the upsampling architecture recovers the information by using the tensor from the final layer of the baseline network and up-scale it. Following that, the feature map will genarate Gaussian Heat Maps based on the Ground truth, as shown in Fig.1. The default heat map dimension is same with the original images 256×192 for images worth 256×192 and 384×288 for images worth 384×288. In order to fix with the resolution of the feature maps throughout the training phase, the heat maps must grasp the image's scale. For regression, the network will utilize the ground truth heat map and these heat maps to generate the predicted human joint.

Attention Module The Attention Mechanism is made up of two primary components, as shown in Fig.2. First, the feature information was sent to the channel attention module following block one in the backbone network (CAM). The feature information in CAM uses global average pooling to reduce the tensors from $W \times H \times C$ to $1 \times 1 \times C$. It first passes through the convolution block, which converts the tensor to $1 \times 1 \times \frac{C}{r}$, where r is the shrinking ratio which is stick to 16. The weight was then triggered by the CAM using the ReLU. The last stage in CAM is to employ a 1x1 convolution layer to resize the channel to $1 \times 1 \times C$ and to normalize the tensor using the sigmoid. The information for CAM were then combined using element-wise multiplication.

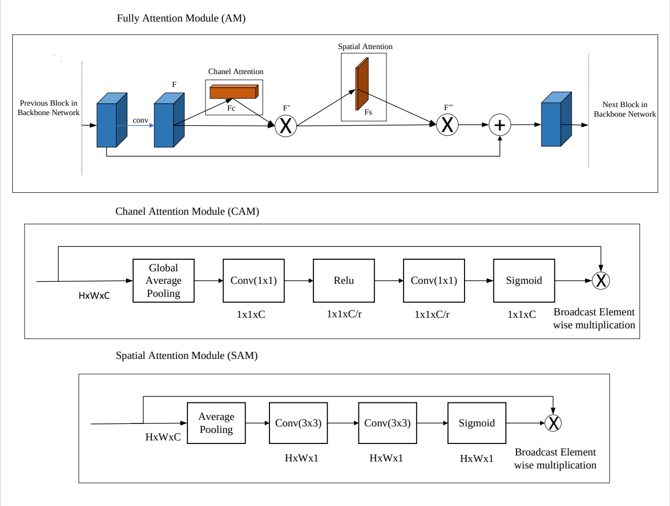
Fully Attention Module (AM)
Previous Block in Backbone Network
conv
F
Chanel Attention
Fc
F'
Spatial Attention
Fs
F"
Next Block in Backbone Network
Chanel Attention Module (CAM)
HxWxC
Global Average Pooling
Conv(1x1)
Relu
Conv(1x1)
Sigmoid
1x1xC
1x1xC/r
1x1xC/r
1x1xC
Broadcast Element wise multiplication
Spatial Attention Module (SAM)
HxWxC
Average Pooling
Conv(3x3)
Conv(3x3)
Sigmoid
HxWx1
HxWx1
HxWx1
Broadcast Element wise multiplication

Fig. 2. Spatial Attention Module (SAM) and Channel Attention Module (CAM) Architecture . At comparison, this picture depicts the description of the attention module, which includes the spatial and the channel module in the bottom and center of the list, respectively, and the whole attention module at the top.

The tensor will be supplied into the Spatial Attention Module after passing through the CAM. The tensors in SAM takes the average pooling for the channel from $W \times H \times C$ to $W \times H \times 1$. Following pooling, convolution layers with kernel size 3×3 were utilized two times to extract the geographical data for the architecture, and the final step in SAM is fed to the CAM shown in Figure 2. Finally, the intended solution employed element-wise extensions to the original tensor and the tensor after AT to be merged, as well as a new tensor for the continuous backbone network block.

B. Loss Function

Heat maps are used in this work to illustrate body joint locations for the loss function. As the ground-truth position in Fig. 1 by $m = \{mj\} J = 1^J$, where $Xj = (xj, yj)$ is the geographical harmonize of the jth body joint for each image. The value of heat map for Ground-truth Hj is then constructed using the Gaussian distribution and the mean aj with variance $\sum$ as shown below.

$$H_j(p) \sim N(a_j, \sum) \quad (1)$$

where $\mathbf{p} \in \mathbb{R}^2$ demonstrate the coordinate, and $\sum$ is experimentally decided as an identity matrix $\mathbf{I}$. The last layer of the neural architecture forecast J heat maps, *i.e.,* $\hat{S} =$

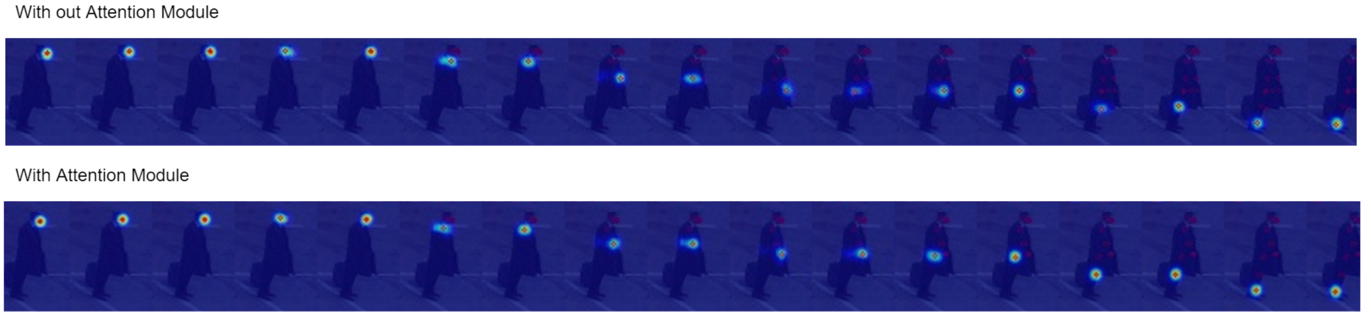

Fig. 3. Predicted Heat-map before and after used Attention Module that includes 17 sequences images for 17 keypoint in COCO dataset

$\left\{\hat{S}j\right\}j = 1^J$ for J body joints. A loss function is defined by the mean square error, which is calculated as follows:

$$L = \frac{1}{MJ}\sum_{m=1}^{M}\sum_{j=1}^{J}\left\|S_j - \hat{S}_j\right\|^2 \tag{2}$$

M denotes the number of selected in the training process. Using data from the last layer or backbone architecture, the trained network generated predict heat maps using groundtruth heat maps.

IV. Experiments

A. Experiment Setup

Dataset. The proposed technique uses the Microsoft COCO 2017 dataset [25] throughout the training and inference process. This dataset comprises around 200K pictures and 250K human samples, each with 17 keypoint labels. The study's data collection includes three folders: train set for training, validation set and test-dev set for testing. Furthermore, the annotations files for train and validate are open to the public and are accompanied by the individualist.

Evaluation metrics. This paper utilized Object Keypoint Similarity (OKS) for COCO [25] with $OKS = \frac{\sum_i exp(-{d_i}^2/2s^2k_i^2)\delta(v_i>0)}{\sum_i \delta(v_i>0)}$ In this case, d_i is the Euclidean distance between the predicted keypoint and the groundtruth, v_i is the target's visibility flag, s is the object scale, and k_i is a joint for seventeen join in COCO 2017 dataset. The standard average accuracy and recall value are then computed. AP and AR are the averages from OKS=0.5 to OKS=0.95, with AP^M representing medium objects and AP^L representing large objects in Table I.

Implementation details The suggested technique employed data increase in model training, such as flip, 40 degrees by outline for rotaion, and scale, which put 0.3 for the factor. For training images, the batch size was stick to 4 and utilize the shuffle function. The total number of epochs in our experiment is 210, with the baseline learning-rate set at 0.001 and multiplied by 0.1 (learning decade factor) at the 170-th and 200-th epoch. The Adam optimizer [26] and the momentum is 0.9 was employed.

All proposed research are carried out using the Pytorch framework and tested on two datasets. The picture input resolution was reduced to 256x192. The model was trained using CUDA 10.2 and CuDNN 7.3 on a single NVIDIA GTX 1080Ti GPU.

B. Experiment Result

TABLE I
The result for apply the attention module for each stage of HRNet

Backbone	Stage	#Param	AP
HRNet-W32	-	28.5M	74.4
HRNet-W32	1	30.2M	75.5
HRNet-W32	1+2	32.9M	76.0
HRNet-W32	1+2+3	36.4M	76.1

The suggested technique compares each circumstance while adding the attention module for each step from stage 1 to stage 3, as shown in Table 1. The Average Precision (AP) demonstrates that using AM in the first stage gains 1.1 in mAP, which boosts accuracy more than using AM in the second and third stages. Furthermore, the AP is enhanced by 1.5 percent, 2.2 percent, and 2.7 percent, respectively, while the number of parameters grows by 5.96 percent, 15.4 percent, and 27.7 percent for adding AM with stages 1, 2, and 3. In our proposed network, we used only 2 blocks of AM in stage 1, 3 blocks for stage 2 and 4 blocks for stage 3.

TABLE II
The result for apply the attention module for each sub-network of HRNet

Backbone	Sub-network	#Param	AP
HRNet-W32	-	28.5M	74.4
HRNet-W32	1	31.1M	75.4
HRNet-W32	1+2	33.8M	75.9
HRNet-W32	1+2+3	35.5M	76
HRNet-W32	1+2+3+4	36.4M	76.1

As shown in Table 2, the proposed approach compares each case while adding the attention module for each step from sub-network 1 to sub-network 4. The Average Precision (AP) shows that utilizing AM in the first sub-network results in a 1.0 increase in mAP, which improves accuracy more than using AM in the second, third, and fourth sub-networks. Furthermore, the AP increases by 1.3 percent, 2.0 percent, 2.6 percent, and 2.7 percent, respectively, while the number of parameters increases by 9.1 percent, 18.6 percent, 24.5

TABLE III
Comparison on Microsoft COCO 2017 Validation Dataset. AM is mean attention module

Method	Backbone	Input size	#Params	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
8-Stage Hourglass [9]	8-Stage Hourglass	256×192	25.1M	66.9	-	-	-	-	-
Mask-RCNN [27]	ResNet-50-FPN	256×192	-	63.1	87.3	68.7	57.8	71.4	-
SimpleBaseline [10]	ResNet-50	256×192	34.0M	70.4	88.6	78.3	67.1	77.2	76.3
SimpleBaseline [10]	ResNet-101	256×192	53.0M	71.4	89.3	79.3	68.1	78.1	77.1
SimpleBaseline [10]	ResNet-152	256×192	68.6M	73.7	91.9	81.1	70.3	80.0	79.0
Fine-tuning AM [16]	ResNet-50	256×192	31.2M	71.4	91.6	78.6	68.2	75.7	76.3
Fine-tuning AM [16]	ResNet-101	256×192	50.2M	72.3	92.0	79.4	68.3	77.1	77.1
HRNetBaseline [17]	HRNet-W32	256×192	28.5M	74.4	90.5	81.9	70.8	81.0	79.8
HRNetBaseline [17]	HRNet-W48	256×192	63.6M	75.1	90.6	82.2	71.5	81.8	80.4
HRNet + our AM	HRNet-W32	256×192	36.4M	76.1	91.0	82.7	71.5	82.9	81.2
HRNet + our AM	HRNet-W48	256×192	71.8M	76.4	91.1	83.1	72.2	83.3	81.4

percent, and 27.7 percent when AM with sub-stages 1, 2, 3, and 4 is included. In our suggested network, we employed three blocks of AM in the first sub-network , three blocks in the second sub-network, two blocks in the third sub-network , and one block in the final sub-network. Fig.3 shows the result of how the attention module impacts the heatmap generate, which shows AM gained significantly better performance heatmap prediction for the left wrist and left elbow keypoint in the seventh and eighth pictures. Moreover, the attention module also helps the network get better for other joints.

COCO datasets result Our result was estimate on COCO validation dataset. The AP in the proposed perspective get better than the Basic High-Resolution standard in whole circumstance of 1.7 AP, 1.3 AP in HRNet-32, HRNet-W48, respectively. Furthermore, the average recall (AR) is 1.4 points higher in the case of HRNet-W32 and 1.2 points higher with the situation of HRNet-W48. The visualize result can see in Fig.3 which show that used attention module make the predicted heat map get more accurate. Figure 4 show the qualitative result for the COCO 2017 dataset, which demonstrated attention module increase the result of AP for the medium and large object in 0.7 AP and 1.5 AP respectively.

However, human pose estimation, like many other designs today, has a number of issues that must be addressed. The first issue was that the images had hidden joints that were hard to train and anticipate. Second, low-resolution human photos must be correctly removed for human body joints. Following that are images of crowd scenarios, in which it is frequently difficult to determine all of the locations of the joints for all participants. Finally, there is a scarcity of information on images with incomplete parts for evaluating human postures.

V. Conclusion

This research shows the effect of the attention module on CNNs, with a focus on High-Resolution networks. Furthermore, our work demonstrates that by not increasing the amount of parameters, the attention module utilized has a bigger effect. On the other hand, the Attention Module highlighted the critical feature map rather than the other component. As a result, the network will improve efficiency, notably for various activities in the field of computer vision. Future research will focus on defining specific applications or settings to be included in our study, such as the surveillance system and the 3D human pose estimation. Another challenge is related to the limitations in assessing human exposure, which restricts the network's accuracy.

Acknowledgement

This result was supported by ”Regional Innovation Strategy (RIS)” through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(MOE) (2021RIS-003)

References

[1] X. Yang, M. Wang, and D. Tao, “Person re-identification with metric learning using privileged information,” *CoRR*, vol. abs/1904.05005, 2019. [Online]. Available: http://arxiv.org/abs/1904.05005

[2] W. Li, R. Zhao, and X. Wang, “Human reidentification with transferred metric learning,” in *Asian Conference on Computer Vision (ACCV)*, 11 2012, pp. 31–44.

[3] Z. Hussain, M. Sheng, and W. E. Zhang, “Different approaches for human activity recognition: A survey,” 2019.

[4] E. Kim, S. Helal, and D. Cook, “Human activity recognition and pattern discovery,” *IEEE Pervasive Computing*, vol. 9, no. 1, pp. 48–53, Jan 2010.

[5] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, “Convolutional pose machines,” 2016.

[6] C.-J. Chou, J.-T. Chien, and H.-T. Chen, “Self adversarial training for human pose estimation,” 2017.

[7] C. Chen and D. Ramanan, “3d human pose estimation = 2d pose estimation + matching,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 5759–5767.

[8] S. Li, L. Ke, K. Pratama, Y. Tai, C. Tang, and K. Cheng, “Cascaded deep monocular 3d human pose estimation with evolutionary training data,” *CoRR*, vol. abs/2006.07778, 2020. [Online]. Available: https://arxiv.org/abs/2006.07778

[9] A. Newell, K. Yang, and J. Deng, “Stacked hourglass networks for human pose estimation,” *CoRR*, vol. abs/1603.06937, 2016. [Online]. Available: http://arxiv.org/abs/1603.06937

[10] B. Xiao, H. Wu, and Y. Wei, “Simple baselines for human pose estimation and tracking,” *CoRR*, vol. abs/1804.06208, 2018. [Online]. Available: http://arxiv.org/abs/1804.06208

[11] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2015.

[12] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” 2015.

[13] G. Moon, J. Y. Chang, and K. M. Lee, “Posefix: Model-agnostic general human pose refinement network,” 2018.

[14] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, “Deepercut: A deeper, stronger, and faster multi-person pose estimation model,” 2016.

[15] V. Dumoulin and F. Visin, “A guide to convolution arithmetic for deep learning,” 2016.

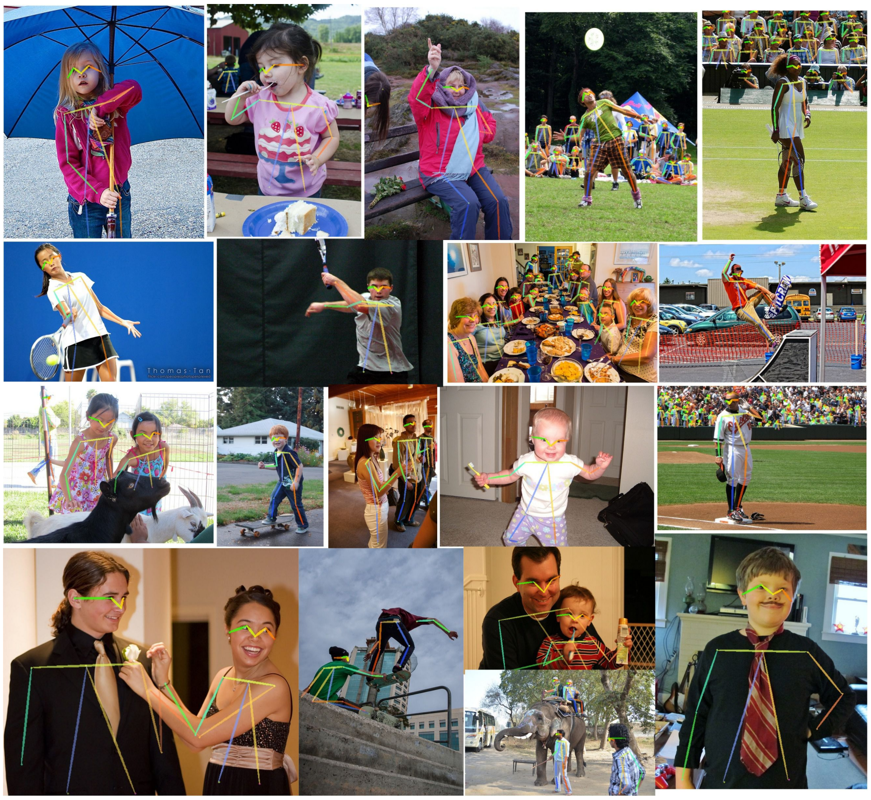

Fig. 4. Qualitative result for human pose estimation in COCO2017 test-dev set

[16] T.-D. Tran, X.-T. Vo, M.-A. Russo, and K.-H. Jo, "Simple fine-tuning attention modules for human pose estimation," in *International Conference on Computational Collective Intelligence*. Springer, 2020, pp. 175–185.

[17] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," 2019.

[18] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," *CoRR*, vol. abs/1312.4659, 2013. [Online]. Available: http://arxiv.org/abs/1312.4659

[19] X. Wang, R. B. Girshick, A. Gupta, and K. He, "Non-local neural networks," *CoRR*, vol. abs/1711.07971, 2017. [Online]. Available: http://arxiv.org/abs/1711.07971

[20] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," 2019.

[21] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," 2017.

[22] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," 2016.

[23] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," 2015.

[24] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," 2018.

[25] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *CoRR*, vol. abs/1405.0312, 2014. [Online]. Available: http://arxiv.org/abs/1405.0312

[26] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 12 2014.

[27] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," 2017.