Efficient High-Resolution Network 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*—Convolutional neural networks (CNNs) have attained the maximum performance today not just for human posture prediction but also for other machine vision applications (e.g., object identification, semantic segmentation, images classification). The Attention Module also reveals their dominance over other traditional networks (AM). As a result, the focus of this research is on creating a suitable feed-forward AM for CNNs. First, input the feature map into the attention module, which is divided into two dimensions: channel and spatial, after passing through a stage in the backbone network. The AM then multiplies these two feature maps and sends them to the backbone's next level. The network may collect more data in terms of long-distance dependencies (channels) and geographic data, resulting in increased precision efficiency. Our findings would also reveal a distinction between the use of the attention module and current approaches. When compared to the baseline-CNN backbone, switching to a High-resolution network (HRNet) keeps the projected joint heatmap accurate while reducing the number of parameters. The suggested architecture outperforms the baseline-HRNet by 2.0 points in terms of AP. The proposed network was also trained using the COCO 2017 benchmarks, which are now available as an open dataset.**

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

I. Introduction

In today's world, 2D human posture estimate plays an essential but difficult role in computer vision, fulfilling a variety of goals such as human re-identification [1], [2], activity recognition [3], [4], human pose estimation [5], [6], and 3D human pose estimation [7], [8]. The fundamental purpose of the human posture is to identify bodily portions for human body keypoints. The importance of channel and spatial backdrop in improving the precision of key point regression cannot be overstated. As a result, the focus of this study will be on how to teach the network to pay better attention to information.

According to recent developments, deep convolutional neural networks have lately achieved outstanding performance. Before raising the resolution, most existing techniques route the input through a network, which is typically made up of high-to-low resolution subnetworks connected in series. Hourglass [9], for example, uses a symmetric low-to-high technique to recover high resolution. SimpleBaseline [10] uses a few transposed convolution layers to build high-resolution representations. Dilated convolutions are also used to increase the last layers of a high-to-low resolution network (such as VGGNet or ResNet) [11], [12].

Deep neural network convolution has now encoded major advancements in human posture [13], [14]. However, these networks face numerous obstacles. To begin, how can the accuracy of various types of networks be improved (For example, a real-time network or a network that measures correctness.) Second, it is common to need to check the speed of a network while updating or modifying it. Finally, the current network must increase accuracy while remaining as fast as possible. This study examines a one-of-a-kind network as well as the speed and accuracy of the attention module. Using and not using the attention module is the subject of the proposed experiment. The experiment also differs from the Simple Baseline [10] experiment in that it does not use the attention mechanism and instead uses the Simple Baseline experiment for upsampling, it instead used the transpose convolution [15]. The proposed method would focus on how productive and economical each network situation is.

The proposed technique was used to create a simple fine-tune attention module called [16], which showed a significant improvement in mean Average Precision (mAP). The proposed network, which is based on VGG16 [11], aims to improve the spatial attention module (SAM) by using two 3×3 convolution layers instead of a 7×7 convolution layer. By employing a 3×3 kernel, the network keeps the mAP while minimizing the implementation cost. In addition, the number of parameters was reduced, which resulted in a faster network. To further comprehend AM, the suggested network increased 4.7 points in Average Accuracy for precision while only increasing 16.5 percent of number parameters, compared to the Attention mechanism standard [16] when using the High-Resolution Network [17] as a backbone network. This study presents a novel network attention module that can quickly respond to a wide range of challenges in a variety of applications, including object recognition, picture classification, and human position estimate. The suggested method uses an up-sampling method to compute joint human posture predictions based on feature map recovery.

II. Related work

2D-Human Pose Estimation: Deeppose [18], the most significant part of human pose estimation is key-point recognition and its interaction with geographical data. Simple baseline employs joint prediction via an end-to-end architecture with a larger constraint. Later, with the Stacked hourglass network

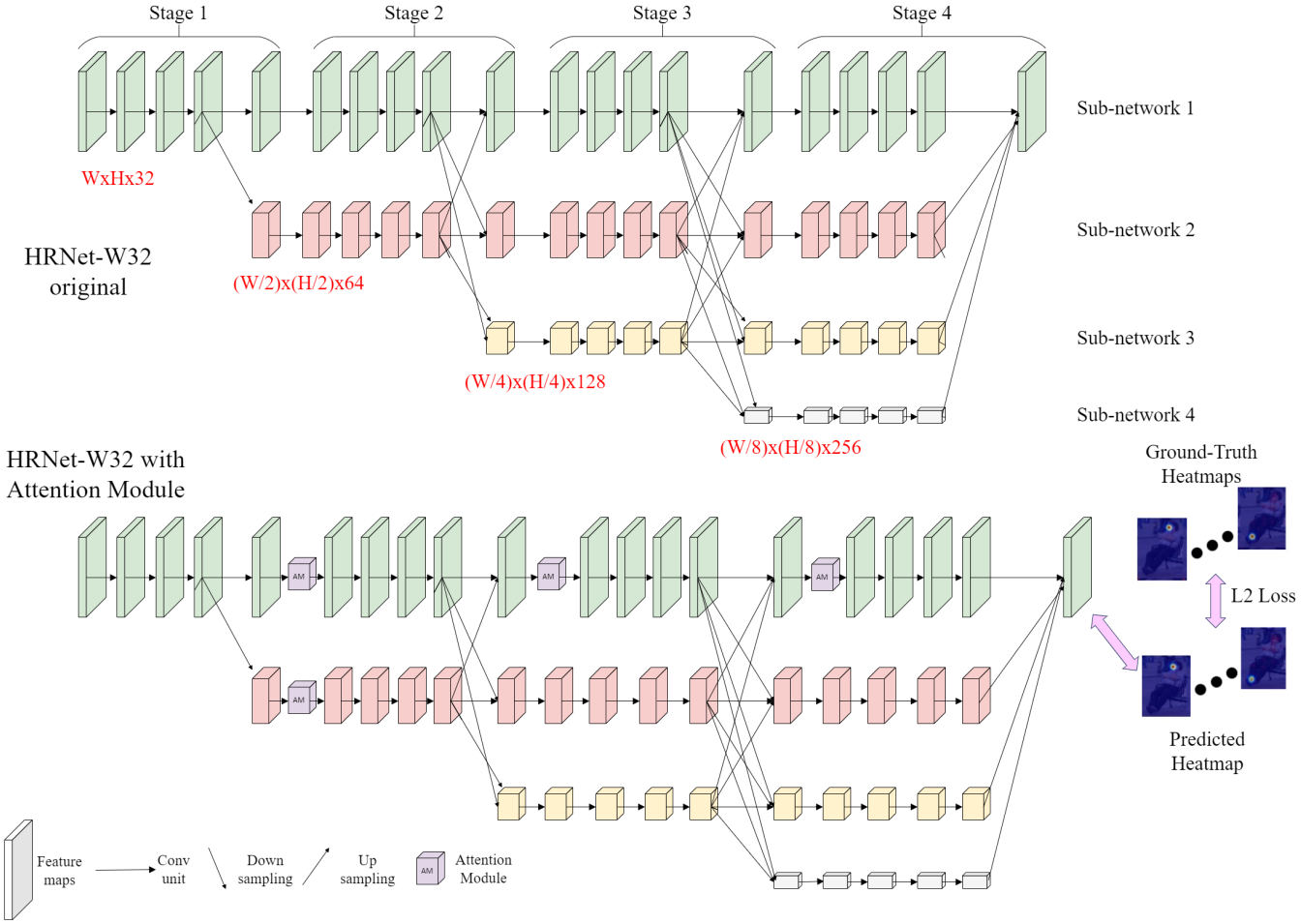

Fig. 1. Illustrate the suggested 2D-human-pose-estimation architecture's outline. The proposed method separated the network into four phases, with an attention module connecting each level.

[9], Newell decreases the number of settings while keeping high accuracy. Nowadays, Sun with the High-Resolution network [17] maintains the high-resolution map from beginning to end to maintain the network's high-level feature till the end. Gaussian distributions were used in all of the approaches to depict local joints. Following that, a convolution neural architecture was used to estimate human posture. To lower employment expenses, they must limit the number of parameters, and utilizing suitable attention tactics will do so. As a result, the proposed technique concentrates on the attention module in use while boosting accuracy and reducing the number of parameters.

A 3×3 kernel size, on the other hand, outperforms a 7×7 kernel size when it comes to enhancing network performance. However, the 7×7 kernel size provides higher precision in certain more complicated and expensive systems. In comparison, our attention module provides an adequate perspective for network design with a limited number of parameters and high speed or a greater number of parameters and lower speed. The essay then shows how the attention module works in each method and consequence.

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.

High Resolution architecture connects high-to-low subnetworks in tandem. It maintains high-resolution representations throughout the process, allowing for geographically accurate heatmap estimation. It generates consistent high-resolution representations by integrating the representations generated by the high-to-low subnetworks periodically. Our method differs from most past efforts in that it necessitates a separate low-to-high upsampling operation as well as an aggregate low-level and high-level feature map. The technique is superior in joint identification accuracy and efficient in computing complexity and parameters without the need for intermediary heatmap monitoring.

Attention mechanism: Human visualization is critical in computer vision, and several focus processing algorithms are being developed to increase CNN efficiency. Wang et al. [19] have proposed a non-local network to collect long-distance dependencies. The SENet Channel Focus Module was merged with the Inception Multi-Branch Convolution in SKNet [20], which was influenced by SENet [21] and Inception citec36. Furthermore, the Module for Geographical Attention is based on Google's STN [22], which collects feature map backdrop data. Additionally, the attention module offers various advantages for saliency detection, multi-label categorization, and individual recognition.

The proposed approach in this study was motivated by the CBAM architecture [23], which uses element-wise multiplication to construct the productive in the midst of both spatial and channel modules. The tensors then add to the preceding tensors to blend the old and new data from the Attention block.

III. Methodology

A. Network architecture

Backbone network: Our system utilized a backbone comprised of HRNet-W32 and HRNet-W48 [17], as depicted in Figure 1 for a complete architecture. Each HRNet is divided into four phases that contain residual blocks and connections. The original RGB image is reduced in size to 256×192 (HRNet-W32, HRNet-W48), the tensor traverses each pillar layer, and the starting resolution of $H \times W$ is reduced twice for each stage. Finally, after traveling down the backbone, the function map's dimension is reduced to $\frac{W}{16} \times \frac{H}{16}$ with 256 channels at the network's final bottom layer. Therefore, the backbone network will only employ the first subnetwork, whose size remains $W \times H$ until the conclusion of the regression. Furthermore, the channels' dimensions were doubled at each level. It progresses from 32 after the first stage to 256 at the end. The baseline network's role is to collect valuable data from extract feature maps and provide it to the Training System, which predicts human joints via cross entropy loss.

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

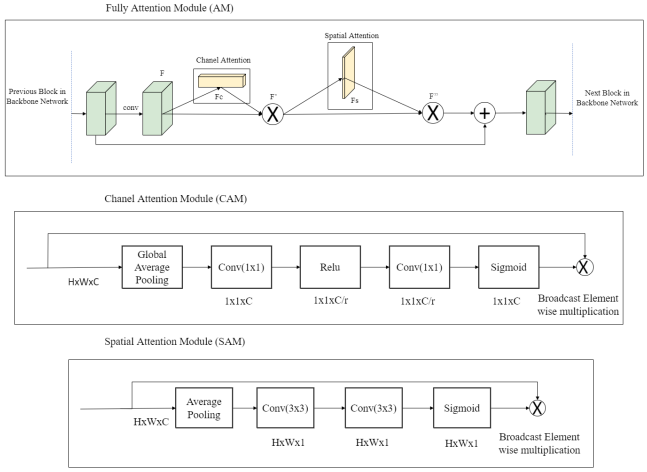

Fig. 2. Architecture of the Spatial Attention Module (SAM) and Channel Attention Module (CAM). This diagram illustrates the attention module's description, which includes the spatial and channel modules at the bottom and center of the list, respectively, and the entire attention module at the top.

$1 \times 1 \times \frac{C}{r}$, 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.

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.

$$H_j(p) \sim N(a_j, \sum) \tag{1}$$

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.,* $\hat{S} = \left\{\hat{S}j\right\} j = 1^J$ 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 ground-truth heat maps.

IV. Experiments

A. Experiment Setup

Dataset. The proposed technique uses the Microsoft COCO 2017 dataset [24] 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 [24] 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 [25] 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

The proposed technique compares each circumstance while adding the attention module for each step from stage 1 to stage 4 (also same with sub-network 1 to sub-network 4) as shown in Table 1. The Average Precision (AP) demonstrates that using AM in the first sub-network and first stage archive 1.3 in mAP, boosts accuracy by 1.75 percent over the baseline. Furthermore, the AP is only decreased from 76.4 to 75.7 compared with the used full of AM for all sub-network and stages while the number of parameters down from 36.4M to 31.9M. In more detail, the number of parameters decreases by 12.4 percent while the accuracy only decreases by 0.9 percent. In our proposed network, we used only 3 blocks of AM in sub-network 1, 1 block for sub-network 2, and total of 4 blocks for the main network.

TABLE I
THE RESULT FOR APPLY THE ATTENTION MODULE FOR EACH SUB-NETWORK AND EACH STAGE 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	76.3
HRNet-W32	1+2+3+4	36.4M	76.4
Backbone	Stage	#Param	AP
HRNet-W32	1	30.2M	75.5
HRNet-W32	1+2	32.9M	76.0
HRNet-W32	1+2+3	36.4M	76.4
HRNet-W32	Sub-1 + Stage-1	31.9M	75.7

COCO datasets result All of our experiment was estimate on COCO validation dataset. The AP in the proposed perspective get better than the Basic High-Resolution standard in whole circumstance of 1.3 AP, 0.9 AP in HRNet-32, HRNet-W48, respectively. Also the average recall (AR) is 1.0 points higher in the case of HRNet-W32 and 0.5 points higher with the situation of HRNet-W48. Furthermore, the AP increase 1.7 percent and AR increase 1.3 percent. Fig. 3 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 0.3 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)

TABLE II
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 [26]	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
Our	HRNet-W32	256×192	31.9M	75.7	90.5	82.2	71.2	82.5	80.8
Our	HRNet-W48	256×192	67.4M	76.0	90.6	82.7	71.8	82.7	80.9

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.

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

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

[24] 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

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

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

Fig. 3. Qualitative result for human pose estimation in COCO2017 test-dev set

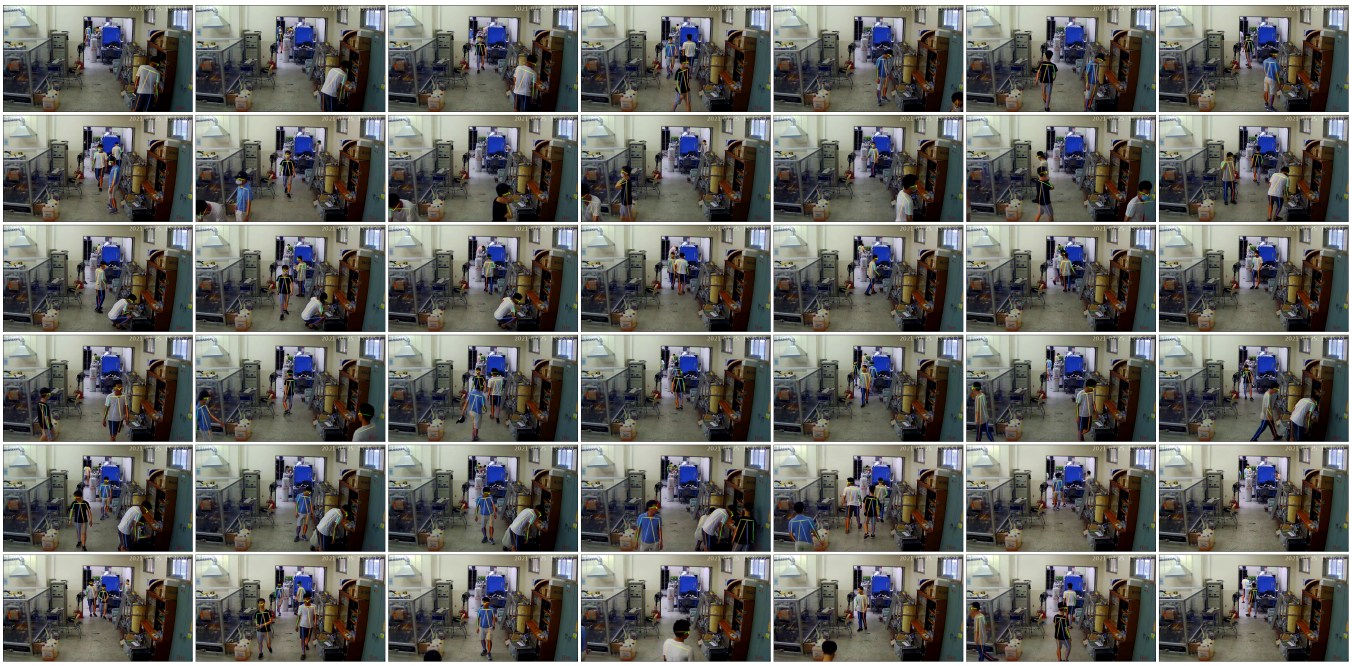

Fig. 4. Qualitative result for human pose estimation tracking in industrial dataset