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*—Convolution neural networks (CNNs) have achieved the best performance nowadays not just for 2D or 3D pose estimation but also for many machine vision applications (e.g., image classification, semantic segmentation, object detection and so on). Beside, The Attention Module also show their leader for improve the accuracy in neural network. Hence, the proposed research is focus on creating a suitable feed-forward AM for CNNs which can save the computational cost also improve the accuracy . First, input the tensor into the attention mechanism, which is divided into two main part: channel attention module and spatial attention module. After that, the tensor passing through a stage in the backbone network. The main mechanism then multiplies these two feature maps and sends them to the next stage of backbone. The network enhance the data in terms of long-distance dependencies (channels) and geographic data. Our proposed research would also reveal a distinction between the use of the attention mechanism and nowadays approaches. The proposed research got better result when compare with the baseline-HRNet by 1.3 points in terms of AP but maintain the number of parameter not change much. Our architecture was trained on the COCO 2017 dataset, which are now available as an open benchmark.**

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

I. Introduction

In today's modern world, 2D human pose estimation plays an essential role and unable to replace in computer vision, which can submit for a variety of goals such as human re-identification, activity recognition, and 3D human pose estimation [1], [2]. The main purpose of the human pose is to identify bodily portions for human keypoints. The importance of channel and spatial backdrop in improving the precision of key point regression cannot be overstated. Hence, the proposed research take focus on how to educate the network to pay better performance to data.

According to recent developments, deep convolutional neural networks have lately achieved outstanding performance. Before raising the resolution, most existing techniques route the input through an architecture, which is typically combine of high-to-low resolution subnetworks connected in series. Hourglass [3], for example, uses a symmetric low-to-high technique to recover high resolution. SimpleBaseline [4] 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) [5], [6].

Deep neural network convolution has now encoded major advancements in human posture [7], [8]. 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 [4] experiment in that it does not use the attention mechanism and instead uses the Simple Baseline experiment for upsampling, the transpose convolution is utilized instead [9].

The proposed research was used to create a simple fine-tune attention mechanism called [10], which showed a significant better performance in mean Average Precision (mAP). The proposed network, which is based on VGG16 [5], aims to improve the spatial attention mechanism (SAM) by using two convolution layers with the kernel size 3×3 instead of a 7×7 convolution kernel. By employing a 3×3 convolution layer, the architecture keeps the mAP while minimizing the implementation fee. Furthermore, the number of parameters was reduced, which resulted in a faster network. To further comprehend AM, the suggested network increased 1.3 points in Average Accuracy for precision while only increasing 11.9 percent of number parameters, compared to the Attention module standard [10] when using the backbone is High-Resolution Network [11]. The proposed study presents a novel attention mechanism that help quickly respond to a wide range of challenges in a variety of applications, including 3D pose estimation, image classification, and semantic segmantation. 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 [12], the most significant part of 2D pose is human joint recognition and its relationship with geographical information. Simple benchmark employs keypoint prediction via an end-to-end network with a bigger constraint. Later, the Stacked hourglass architecture

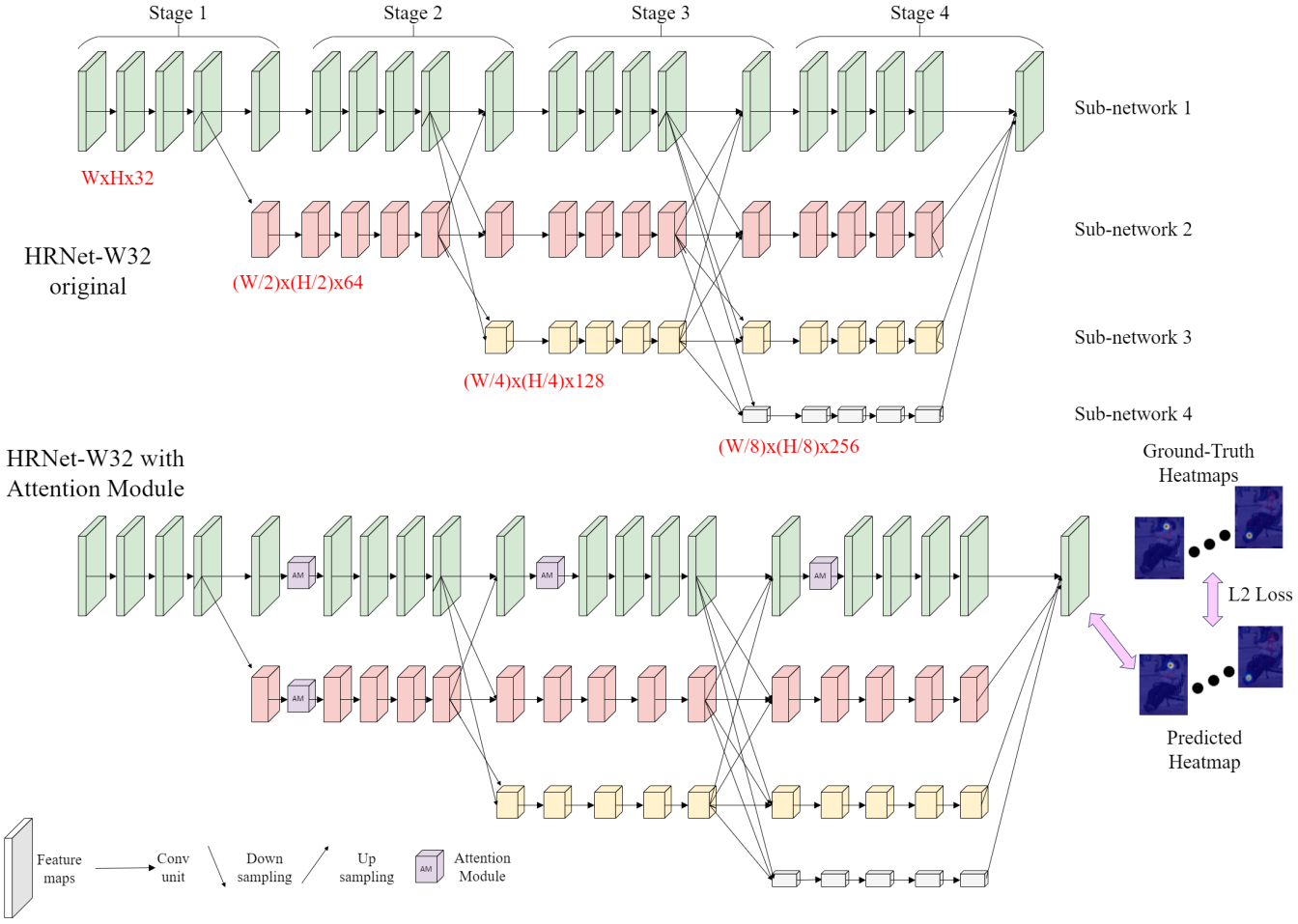

Fig. 1. Illustrate the suggested 2D-pose-estimation outline. The main method separated the architecture into four phases, with an attention module connecting each level.

[3], Newell decreases the number of settings while keeping high accuracy. At present, the High-Resolution network [11] proposed by Sun, keep the high-resolution map from the first stage to the final stage 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 CNN network was used to estimate keypoint detection. To lower employment expenses, the CNN network need limit the number of parameters, and utilizing suitable attention tactics will do so. Follow by the result, the proposed technique concentrates on the attention mechanism in use while boosting precision and reducing the computational cost.

A 3×3 kernel size, on the other hand, improve the result than a kernel with the size of 7×7 when it comes to enhancing network performance. However, the kernel with 7×7 size improves higher precision in certain more complicated and expensive systems. Our attention module, in contrast, offers a suitable viewpoint for network design with a few parameters and high speed or more parameters and slower speed. The attention module's function in each strategy and outcome is then demonstrated in the essay.

High resolution network: The majority of convolution neural networks for joint estimator consist of a regressor that evaluate the heatmaps where the keypoint coordinate are estimated and then convert in default size, a main human pose that representations with the similar dimension as its input, and a stem subnetwork that decreases the resolution. Maintaining the full resolution improves the network's accuracy. High-to-low and low-to-high structures make up the majority of the main body; multi-scale fusion and intermediate (deep) supervision may be included as extras.

In parallel, high-to-low subnetworks are connected via high resolution architecture. Heatmap estimate is made possible by the process' maintenance of high-resolution depiction throughout. By regularly integrating the depiction produced by the high-to-low subnetworks, it produces consistent high-resolution representations. Our approach varies from the majority of prior attempts in that it requires both an aggregate low-level and high-level feature map in addition to a separate low-to-high upsampling procedure. Without the need of a middle heatmap monitor, the approach is more accurate at joint recognition and efficient at computing complexity and parameters.

Attention mechanism: In computer vision, Human Body

visualization is critical and several focus processing algorithms are being developed to increase CNN efficiency. Wang et al. [13] have proposed a non-local architecture to collect long-distance dependencies. The SENet Channel Focus Module was merged with the Inception Multi-Branch Convolution in SKNet [14], which was influenced by SENet [15] and Inception [16]. Furthermore, the Mechanism for Spatial Attention is based on Google's STN [17], which collects feature map backdrop data. Additionally, the attention module offers various advantages for multi-label classification, salience detection and individual recognition.

The proposed architecture in this study was inspired by the CBAM network [18], which uses element-wise multiplication to construct the productive in the midst of both channel and spatial attention mechanism. The tensors then add to the preceding tensors to blend the new and old information from the Attention mechanism.

III. Methodology

A. Network architecture

Backbone network: Our system utilized a benchmark comprised of HighResolutionNet-W32 and HighResolutionNet-W48 [11], as depicted in Fig. 1 for a complete network. Each HighResolutionNet is divided into four aspect that contain skip connections and residual blocks. The default data image is reduced in dimension to 256×192 (HighResolutionNet-W32, HighResolutionNet-W48), the information of feature traverses each pillar layer, and the starting dimension of $H \times W$ is reduced twice for each stage. After the tensor traveling down the backbone, the tensor map's size is reduced to $\frac{W}{16} \times \frac{H}{16}$ with 256 channels at the network's final layer. Therefore, the backbone architecture will only employ the first subnet, whose dimension remains $W \times H$ until the conclusion of the regression. Additionally, the channels' dimensions were increase 2 times at each level. The tensor channel increase from 32 at the first stage to 256 at the end. The baseline architecture's role is to collect valuable information from extract tensor and provide it to the Training process, which predicts human joints via cross entropy loss.

The upsampling network fix the data by utilized the tensor from the last layer of the baseline network and up-scales it after removing the useful data from the backbone architecture. The original heat map size is identical to the original photos for images valued at 256×192 and 384×288. Aiming to match the dimension of the tensor during training, the heat maps must comprehend the size of the picture. The ground truth heat map and both of these heat maps will be used by the network to predict the human joint.

Attention Module According to Fig. 2, the Attention Mechanism consists of two main parts. After block one in the backbone network, the feature information was first transferred to the channel attention module (CAM). The tensor information in CAM utilize a GAP layer (Global average Pooling) to decrease the tensors from $W \times H \times C$ to $1 \times 1 \times C$. Then, It traverse through the convolution layer, which made the important feature into $1 \times 1 \times \frac{C}{r}$, where the shrinking ratio

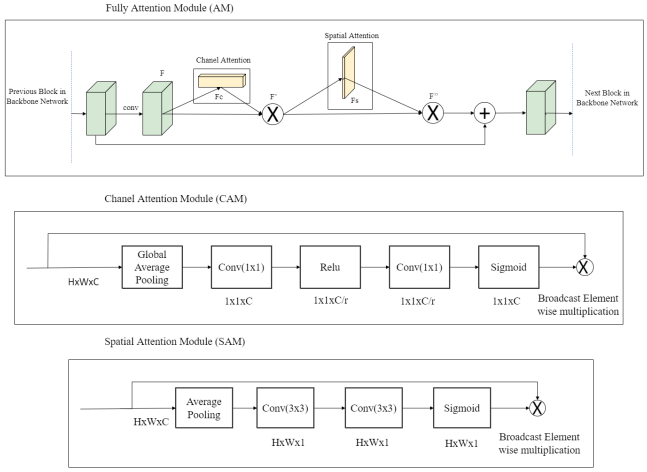

Fig. 2. Architecture of the SAM Module and CAM Module. This diagram illustrates the attention mechanism, which conduct the channel and spatial part at the center and bottom of the figure, respectively, and the entire attention mechanism on the top.

is r which set at default to 16. The weight inside network was triggered by the Channel mechanism utilizing function ReLU for the activation. Finally, the proposed CAM utilize a 1x1 convolution layer to made the size of channel to $1 \times 1 \times C$ and apply sigmoid to normalize the weight in final tensor. The data in CAM were then mixed by utilize a multiplication of element-wise.

After going through the CAM, the tensor will be sent into the SAM. The tensors in Spatial module change the channel's by apply average pooling so the tensor from $W \times H \times C$ to $W \times H \times 1$. The last step in SAM is supplied to the CAM depicted in Figure 2 after pooling, and convolution layers with 3×3 kernel size were used twice to extract the geographic data for the network. The suggested solution also included a new tensor for the continuous backbone architecture block and element-wise additions to the default tensor and the tensor after AT to be combined.

B. Loss Function

Heat maps are utilize in the proposed work to demonstrate body keypoint locations in the loss function. At the beginning, we set the ground-truth point by $m = \{m_n\}\, N = 1^N$, where $X_n = (x_n, y_n)$ is the geographical information of the n^{th} body keypoint for every image. The principles of Ground-truth heat map H_n is then build up by utilized the Gaussian distribution and the mean a_n with variance $\sum$ as illustrate in the next equation.

$$H_n(p) \sim N(a_n, \sum) \tag{1}$$

where $\mathbf{p} \in \mathbb{R}^2$ illustrate the coordinate, and $\sum$ is automatically decided as an identity matrix $\mathbf{I}$. The final layer of the proposed architecture generated J heat maps, *i.e.,* $\hat{S} = \left\{\hat{S}n\right\} n = 1^N$

for K body joints. Mean square error for loss function is defined, which is summarize as follows:

$$L = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} \left\| S_n - \hat{S}_n \right\|^2 \quad (2)$$

M illustrate the number of images was selected in the training part. Utilizing the information from the final layer or the proposed architecture, the trained model created the predicted heat maps by used ground-truth one.

IV. Experiments

A. Experiment Setup

Dataset. Microsoft COCO 2017 dataset [19] was utilized through the training and testing process. This dataset is a challenge dataset for joint detection which comprises around 250K human labelled in 200K images, each human pose have 17 keypoint labels. The benchmark dataset divide the data into three folders: train2017 for training process, val2017 for validation test and test-dev for testing. Additionally, the individualist is included in the publically accessible annotation files for train and validate.

Evaluation metrics. The proposed research apply Object Keypoint Similarity (OKS) for Microsoft COCO2017 dataset with $OKS = \frac{\sum_i exp(-{d_i}^2/2s^2k_i^2)\delta(v_i>0)}{\sum_i \delta(v_i>0)}$ In the above function, The Euclidean distance between the groundtruth joint and the predicted joint is d_i , The target's visibility flag is v_i, The object scale is s, and k_i is one of seventeen joints in Microsoft COCO 2017 Benchmark. Hence, The standard average accuracy and recall value are then computed. In addition, AP (Average Precision) and AR (Average Recall) are the averages from OKS=0.5 to OKS=0.95, with AP^L mean AP for large objects and AP^M mean AP for medium objects in Table II .

Implementation details The recommended method made use of an increase in data during model training for operations including flip, scale, and rotation at an angle of 40 degrees relative to an outline. The batch size for training photos was kept at 4 and the shuffle function was used. In our experiment, there are 200 total epochs. The learning rate default is set 0.001 and is doubled by 0.1 (learning decade factor) at the 160-th and 190-th epochs. The momentum was 0.9 and the Adam optimizer was used.

The Pytorch framework is used for the suggested research, which is evaluated on two set. The supplied image dimension was scaled down to 256x192. On one NVIDIA GTX 1080Ti GPU, the architecture trained by utilized CuDNN 7.3 and CUDA 10.2.

B. Experiment Result

In Table 1, The proposed research contrast each parts of network while adding the attention mechanism for each part from stage 1 to stage 4 (also same with first sub-network to final sub-network). The AP illustrates that utilize Attention mechanism in the first sub-network and first stage archive 1.3 in mAP, improves 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 the proposed architecture, we utilized only 3 AM blocks in sub-network 1, 1 block for sub-network 2, and total of 4 blocks for the main network.

TABLE I
THE RESULT WHEN UTILIZE THE ATTENTION MECHANISM FOR EACH SUB-NETWORK AND EACH STAGE OF HIGH RESOLUTION NETWORK

Backbone	Sub-Net	AP	#Param
HighResolutionNet-W32	-	74.4	28.5M
HighResolutionNet-W32	1	75.4	31.1M
HighResolutionNet-W32	2+1	75.9	33.8M
HighResolutionNet-W32	3+2+1	76.3	35.5M
HighResolutionNet-W32	4+3+2+1	76.4	36.4M
Backbone	Stage	#Param	AP
HighResolutionNet-W32	1	75.5	30.2M
HighResolutionNet-W32	2+1	76.0	32.9M
HighResolutionNet-W32	3+2+1	76.4	36.4M
HighResolutionNet-W32	Sub-1 + Stage-1	75.7	31.9M

COCO datasets result All of our experiment was experiment on COCO2017 validation set. The Average Precision(AP) in the main network got better performance than the original High-Resolution in whole case of 1.3 AP, 0.9 AP in HighResolutionNet-32, HighResolutionNet-W48, respectively. Also the average recall (AR) is 1.0 points better in the situation of HighResolutionNet-W32 and 0.5 points better with the case of HighResolutionNet-W48. Furthermore, the AP increase 1.7 percent and AR increase 1.3 percent. Fig. 3 illustrate the qualitative 2D pose for the Microsoft COCO2017 benchmark, which show attention mechanism improve the final result of AP for the large object and medium object in 1.5 AP and 0.3 AP respectively.

2D pose estimation, similar with other modern architecture, has a variety of problems that need to be solved. The pictures' concealed keypoint, which were challenging to train for and predict, were the first problem. Second, joints in the human body must be accurately eliminated from low-resolution human images. The photographs that follow show crowd situations, when it is usually challenging to pinpoint where each participant's keypoint are located. Last but not least, there is a dearth of data on photos with missing pieces for assessing human pose case.

V. Conclusion

The attention module's impact on CNNs is demonstrated in this study, with a particular emphasis on high-resolution networks. Additionally, our research shows that the attention module is more effective when fewer parameters are used. The Attention Module, however, chose to emphasize the key tensor rather than the other parts of network. Follow by the result, the proposed architecture operate more effectively, especially for various computer vision-related tasks. The definition of certain settings or applications to be included in our research, such

TABLE II
COMPARISON RESULT ON COCO 2017 VALIDATION SET.

Methodology	Backbone	#Parameters	Image dimension	AP	AR	AP^{50}	AP^{75}	AP^{L}	AP^{M}
Hourglass [3]	8 Hourglass network	25.1M	256×192	66.9	-	-	-	-	-
Mask-RCNN [20]	ResNet-50-FPN	-	256×192	63.1	-	87.3	68.7	71.4	57.8
Benchmark [4]	ResNet-50	34.0M	256×192	70.4	76.3	88.6	78.3	77.2	67.1
Benchmark [4]	ResNet-101	53.0M	256×192	71.4	77.1	89.3	79.3	78.1	68.1
Benchmark [4]	ResNet-152	68.6M	256×192	73.7	79.0	91.9	81.1	80.0	70.3
Fine-tune Attention [10]	ResNet-50	31.2M	256×192	71.4	76.3	91.6	78.6	75.7	68.2
Fine-tune Attention [10]	ResNet-101	50.2M	256×192	72.3	77.1	92.0	79.4	77.1	68.3
High-Resolution Net [11]	HighResolutionNet-W32	28.5M	256×192	74.4	79.8	90.5	81.9	81.0	70.8
High-Resolution Net [11]	HighResolutionNet-W48	63.6M	256×192	75.1	80.4	90.6	82.2	81.8	71.5
Our	HighResolutionNet-W32	31.9M	256×192	75.7	80.8	90.5	82.2	82.5	71.2
Our	HighResolutionNet-W48	67.4M	256×192	76.0	80.9	90.6	82.7	82.7	71.8

as the 3D human pose estimation and inside the surveillance system, will be the subject of future research. The network's accuracy is further limited by the challenges associated in determining human pose.

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] 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.

[2] 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

[3] 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

[4] 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

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

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

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

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

[9] V. Dumoulin and F. Visin, "A guide to convolution arithmetic for deep learning," 2016.

[10] 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.

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

[12] 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

[13] 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

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

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

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

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

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

[19] 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

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

Fig. 3. Qualitative 2D pose estimation in Microsoft COCO 2017 test set