An Efficient Method for Boosting Human Pose Estimation

Shicheng Xiang, Xiao Chen, Jun Zhou*
Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai, China
Shanghai Key Lab of Digital Media Processing and Transmission
$Email : \{scheng.xiang, zhoujun, autumnrain\}@sjtu.edu.cn$

2021 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB) | 978-1-6654-4908-3/21/$31.00 ©2021 IEEE | DOI: 10.1109/BMSB53066.2021.9547183

***Abstract*—Most existing human pose estimation approaches fall into designing new network architectures or tend to apply deeper layers. Most of the methods are time-consuming, and lightweight plug-in for improving human pose estimation to the existing network gets little investigation. In this paper, we propose a lightweight plug-in module to boost the performance of human pose estimation named PoseReNed. PoseReNet is a network with three branches that are designed to tackle the attenuation caused by occluded keypoints or different scales of the keypoints. Generally, small-scale keypoints are more difficult to detect, and we observe that different channels of the output feature map have different attributes to the performance of estimation. We apply a channel attention mechanism to re-weight the channel to trade-off among different scales of the keypoints. By aggregating multiscale output feature maps, the pose estimation performance can be improved. Serving as a model-agnostic plug-in, PoseReNet brings about significant performance boost to existing human pose estimation models. Extensive experiments show that PoseReNet can effectively improve precision on COCO and MPII.**

***Index Terms*—Human Pose Estimation, Plug-in, Attention mechanism**

I. Introduction

2D Human pose estimation is a challenging but fundamental computer vision task that aims to detect human body key points (e.g., head, wrist, heap, etc.) or anatomical parts coordinate in images [1]. Since the varieties of the appearance of human articular point, occlusion, and complex environmental context, localizing accuracy coordinates of key points is a non-trivial problem. Many practical applications have relation to human poses, such as re-identification, human action recognition, animation [2]–[4] etc. For years, human pose estimation was dependent on handcraft features such as SIFT [5]. Recently, since the excel of deep convolutional neural network (CNN) in this task [6]–[10], existing mainstream approaches of human pose estimation are based on the backbone CNN architectures designing and the state-of-the-art pose estimation methods [8]–[11] are based on the CNN model.

Most existing methods pass the input through a network, and the network can process the input and can be presented as follow:

$$out = network(input) \tag{1}$$

$$coordinates = inference(out) \tag{2}$$

The input is an image after preprocessing, and there are two kinds of representation of out, coordinates and heatmaps. Before the model processing input, the images need to be preprocessed, such as crop, flip and affine transformation, to enhance the robustness of data and thus improving the estimation effects. Additionally, the 2D ground-truth coordinates(x, y) would be translated to heatmap formation(i.e., encoding stage), and it will be used as supervision with predicted heatmap together during the training stage.

Recent research by Moon $et\ al.$ [12] give us a clue that there is another way to boost the performance of human pose estimation, they utilize the error distribution statistics of human pose datasets summarized by Ronchi $et\ al.$ [13] as prior information to generate synthetic poses, then utilize the synthesized poses to train CNN model and estimation results of the model will be input to the proposed method during the testing stage. To train their model, they generate each type of the errors (i.e., jitter, inversion, swap, and miss) based on the pose error distributions to train models, and construct diverse and realistic poses. The generated input pose is fed to the model with the input image, and the model learns to refine the pose [13].

Dark [14] find that the encoding stage(i.e.transforming ground-truth coordinates to heatmaps) and decoding stage can make a huge impact on estimation. They modify the standard coordinate encoding method by generating unbiased heatmaps. Their main contributions fall into two folds: (i) unbiased sub-pixel centred coordinate encoding and (ii) an efficient coordinate decoding method based on Taylor-expansion. They explored and used the deficiency within coordinate encoding and decoding.

In this work, motivated by [11], [13], we propose a new method named PoseReNet to refine human pose estimation instead of designing a complex and deep backbone network. Our contributions are three folds: (i) We design a channel attention mechanism to reweight heatmaps channels, for different channels of heatmaps represent different scales of keypoints. It solves the problem that small-scale keypoints are more difficult to detect. (ii) We apply spatial attention mechanism to fuse global context information, which is helpful for global keypoints inference. It can be helpful to inference occluded keypoints. (iii) We integrate context information and spatial information by aggregating output heatmaps.

II. Related Work

A. Single-person pose estimation.

2D single-person pose estimation is to locate the body joint position of a single person in the input image. For images with multiple persons, preprocessing is needed to crop the original image, such as using a human detector, so that there is only a single person in the input image. The main methods include the early method of directly returning the key point coordinates and the later method of determining the key point coordinates by predicting the heatmap.

DeepPose [6] is one of the earliest methods that apply deep learning for human pose estimation. DeepPose has made a pioneering contribution in the field of human pose estimation, which has greatly promoted the development of deep learning in the task of human pose estimation.

CPM (convolutional pose machines) [15] method studies the dependence between keypoints. The idea is to learn the expression of spatial information through convolutional networks, use different sizes of receptive fields to deal with the scale problems of key parts, and use large receptive fields to infer the occluded joint points.

Stacked Hourglass Network [10] can better obtain the information of the key points of different scales in the image, and combine the information of these different scales so that the network achieved better performance for the overall human target in the image. This method can better solve the problem of target scale change caused by the distance of the viewpoint.

As Tang *et al.* proposed in [16] that most of the previous human pose estimation network models shared the extracted features, which is unreasonable. Therefore, they propose a part-based branching network(PBN) similar to DLCM [17]. PBN groups all related nodes, but its grouping method is obtained by calculating the household information between the joint points in the human body pose estimation data set.

B. Multi-person pose estimation.

Different from single-person pose estimation, multi-person key points estimation has deals with detection and grouping tasks, because there is no prompt for how many people are in the input image. According to which level (high-level abstraction or low-level pixel evidence) the calculation starts, the human pose estimation method can be divided into Top-down and Bottom-up methods.

Models such as Hourglass [10], CPN [18], and Simple Baseline [9] are a process from high resolution to low resolution and then from low resolution to high resolution when extracting features. Sun *et al.* [8] believe that this approach may lose high-resolution information. They proposed a high-resolution net (HRNet). Different from other methods that use low-resolution feature maps to restore high-resolution feature maps, HRNet added the extraction and retention of high-resolution feature maps, so that when the network performs low-resolution feature extraction, it can perform higher resolution in parallel feature extraction.

Newell *et al.* [19] designed an end-to-end bottom-up architecture based on Hourglass Network, and a new Associative Embedding method to supervise the output feature map. This method used the modified Hourglass Network to predict the heatmap of all human joint points in the image, and at the same time generates a corresponding label heatmap for joint point grouping.

III. PoseReNet

In view of the existing research, there is less attention to improving the network, and the existing difficulties for human pose estimation are mainly the occlusion of key points and the inconsistency of scales, to deal with these problems, we propose a lightweight refine model named PoseReNet, which can servers as a plug-in for the current backbone network. The overall pipeline of our methods is shown in Fig.1. The number of channels of each feature block equals the number of key points. The loss in Fig.1 is $l2$ loss:

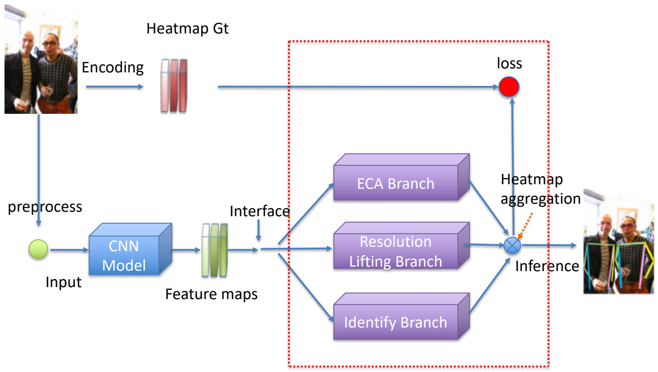


Fig. 1: The pipeline of our human pose estimation system

$$loss(x, y) = \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i)) \tag{3}$$

A. Channel attention mechanism branch

In recent years, the introduction of channel attention into convolutional blocks has attracted widespread attention and has shown great potential in performance improvement. One of the representative methods is ECA-Net [20], which can learn the channel attention of each convolution block and bring significant performance gains to various deep CNN architectures.

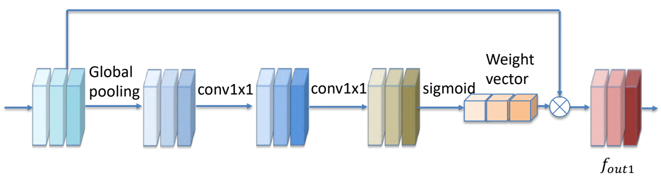


Fig. 2: ECA Branch of PoseReNet

We propose a lightweight but effective channel attention (ECA) branch that corresponds to branch 1 in Fig.1, which only adds a small number of parameters, but can obtain significant performance gains. Then ECA Branch can be formulated as Equation.4.

$$f_{out1} = sigmoid(conv_{1\times1}(conv_{1\times1}(Gp(f_{in})))) \times f_{in} \quad (4)$$

where $GP()$ is globalpooling.

Since different key points in the human body keep different scales, the small key points are more difficult to detect than large key points (e,g., the wrist is smaller than the heap, and therefore wrist is more difficult to locate than heap). To tackle such a problem, we design a weight vector to reweight features in different channels.

B. Resolution lifting branch

The second branch is for generating higher resolution feature maps which is two times higher than the input feature maps. We apply a transposed convolution operation to upsample the input feature maps, and then apply ground truth coordinates to generate a high resolution feature map as supervision. High resolution supervision can be served as a spatial attention mechanism to fuse global context which is beneficial to global inference, and global context information can help to improve key points estimation when key points occluded. The details of the resolution lifting branch is shown in Fig.3. It can be described by:

$$f_{out2} = sigmoid(conv_{3\times3}(conv_{1\times1}(Decove(f_{in})))) \quad (5)$$

where $Deconve()$ is transposed convolution operation.

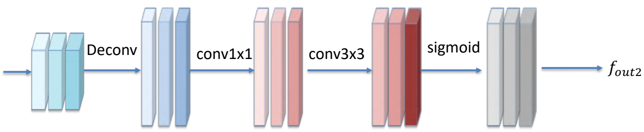


Fig. 3: Resolution lifting branch of PoseReNet

C. Identify branch

The third branch of PoseReNet is an identify branch, the purpose of this branch is to keep the original information. It can be described by the following formula6.

$$f_{out3} = conv_{3\times3}(conv_{3\times3}(f_{in})) \quad (6)$$

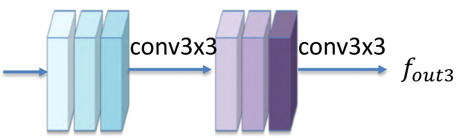


Fig. 4: Identify branch of PoseReNet

Fig. 5: Results of estimation on COCO. The first row is estimated on single person images, the second and third row are estimated on complicated scenes such as multi-persons, occluded and different scales of persons

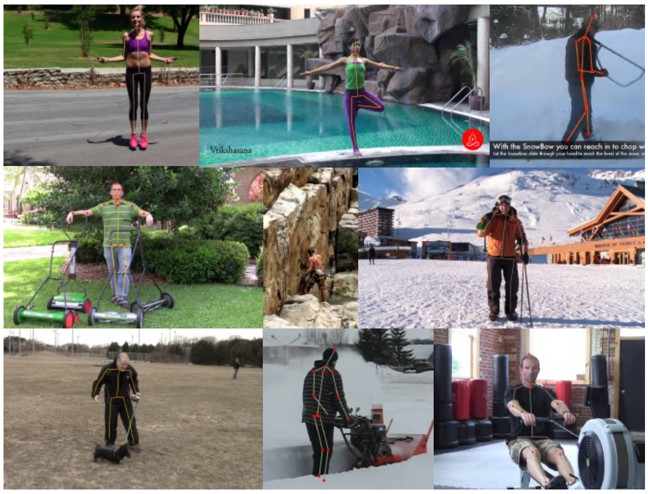


Fig. 6: Results of estimation on MPII

Finally, we aggregation heatmaps of the three branches and the PoseReNet module can be described as Equation.7.

$$f_{out} = Aggregation(f_{out1}, f_{out2}, f_{out3}) \quad (7)$$

IV. Experiments

A. Datasets and Evaluation metric

We do experiments on two mainstream datasets, COCO and MPII. The COCO key points dataset [24] includes over 200,000 images and 250,000 person instances labeled with 17 keypoints. COCO dataset was divided into *test/train/test-dev* subset. Our experiments are all trained on *train* subset. The generic evaluation metric of COCO are OKS, AP and AR.

$$OKS = \frac{\sum_i exp(-d_i^2/2s^2k_i^2)\delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \quad (8)$$

978-1-6654-4908-3/21/$31.00 ©2021 IEEE

TABLE I: Performance Comparison on COCO dataset. "**-**" means original backbone method, "**+**" means adding PoseReNet to backbone method. $AP^{0.5}$ means AP at $OKS = 0.5$, AP^M means AP for medium objects, AP^L means AP for large objects

PoseReNet	Backbone	Input size	#Params(M)	GFLOPs	AP	AP^{50}	AP^{75}	AP^M	AP^L	AR
-	SimpleBaseline-R50+Dark	128×96	34.0	2.3	62.6	86.1	70.4	60.4	**67.9**	69.5
+					**63.1**	**88.3**	**70.6**	**61.6**	65.8	**70.1**
-	SimpleBaseline-R101+Dark	128×96	53.0	3.1	63.2	86.2	71.1	61.2	68.5	70.0
+					**63.9**	**89.3**	**71.5**	**62.4**	**68.7**	**70.2**
-	HRNet-W32+Dark	128×96	28.5	1.8	70.7	88.9	78.4	67.9	**76.6**	**76.7**
+					**71.8**	**89.7**	**79.3**	**69.8**	75.3	75.0
-	HRNet-W48+Dark	128×96	63.6	3.6	71.9	89.1	79.6	69.2	**78.0**	77.9
+					**73.5**	**92.5**	**81.5**	**71.4**	76.6	**78.3**
-	HRNet-W32+Dark	256×192	28.5	7.1	75.6	90.5	82.1	71.8	**82.8**	80.8
+					**76.2**	**90.6**	**82.7**	**73.2**	82.0	**80.9**

TABLE II: Performanc Comparsion with mainstream and start-of-the-art methods on the COCO dataset.

Method	Backbone	Input size	#Params(M)	GFLOPs	AP	AP^{50}	AP^{75}	AP^M	AP^L	AR
G-RMI [21]	ResNet-101	253×257	42.6	57.0	64.9	85.5	71.3	62.3	70.0	69.7
CPN [18]	ResNet-Inception	384×288	-	-	72.1	91.4	80.0	68.7	77.2	78.5
RMPE [22]	PyraNet	320×256	28.1	26.7	72.3	89.2	79.1	68.0	78.6	-
CFN [23]	-	-	-	-	72.6	86.1	69.7	78.3	64.1	-
CPN(ensemble) [18]	ResNet-Inception	384×288	-	-	73.0	91.7	80.9	69.5	78.1	79.0
SimpleBaseline [9]	ResNet-152	384×288	68.6	35.6	73.7	91.9	81.1	70.3	80.0	79.0
HRNet [8]	HRNet-W32	384×288	28.5	16.0	74.9	92.5	82.8	71.3	80.9	80.1
HRNet [8]	HRNet-W48	384×288	68.6	32.9	75.5	92.5	83.3	71.9	81.5	80.5
Dark [14]	HRNet-W32	128×96	28.5	1.8	70.0	90.9	78.5	67.4	75.0	75.9
Dark [14]	HRNet-W48	384×288	63.6	32.9	76.2	92.5	83.6	72.5	82.4	81.1
Ours(Dark+PoseReNet)	HRNet-W32	128×96	23.5	1.8	71.8	89.7	79.3	69.8	75.3	75.0
Ours(Dark+PoseReNet)	HRNet-W32	256×192	28.5	7.1	76.2	90.6	82.7	73.2	82.0	80.9
Ours(Dark+PoseReNet)	HRNet-W48	384×288	63.6	32.9	**77.3**	**92.6**	**83.8**	**74.1**	**85.2**	**81.2**

TABLE III: Comparsion with the start-of-the-art methods on the MPII dataset.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Mean
PCKh@0.5								
HRNet-W32	**97.1**	95.9	90.3	**86.5**	89.1	**87.1**	83.3	90.3
Ours(HRNet-W32+PoseReNet)	**97.1**	**96.1**	**90.8**	86.2	**89.4**	86.7	**83.4**	**90.4**

where d_i is the Euclidean distance between the ground truth coordinates and the detected coordinates. v_i denotes the visibility(0 means occlude, 1 means visible) of key points. k_i is a constant that controls falloff and s denote scale. We report standard average precision and recall scores.

$$Precision = \frac{tp}{tp + fp} \tag{9}$$

tp means true positive, fp means false positive,fn means false nagetives.AP means average precision, AR means average recall.

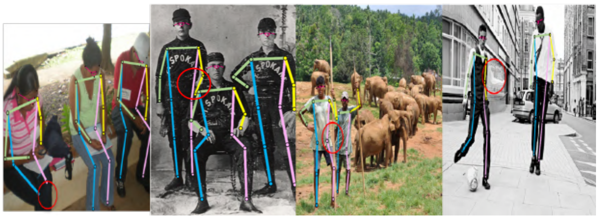
$$Recall = \frac{tp}{tp + fn} \tag{10}$$

The MPII dataset [1] includes 40k person images, each person labeled with 16 joints. Our split method followed the standard *train/val/test* split method [1]. We use PCK (Percentage of Correct Key points)for MPII to evaluate the performance of our method.

B. Implementation details

We adopted the Adam optimizer for model training. Since our method servers as a plug-in, there is no need to modify the original backbone model or just need light modification, we followed the default epochs and learning schedule as in HRNet [8] or SimpleBaseline [9]. For Hourglass, the initial learning rate was set to $2.5e-4$, and declined to $2.5e-5$ and $2.5e-6$ at the 90-th and 120-th epoch, respectively. We adopt three different sizes as input in our experiments. Following [19], we use data augmentation with affine transformation, random rotation($[-30°, 30°]$), random scale ($[0.75, 1.25]$), and random flip as well as random translation to crop input images. Like [8], half body data augmentation is also applied.

978-1-6654-4908-3/21/$31.00 ©2021 IEEE



(a) Results of HRNet on COCO

(b) Results of HRNet+PoseReNet on COCO

Fig. 7: Results comparsion between HRNet (a) and our method (b)

C. Experiment results

We compared our method with the current mainstream and start-of-the-art methods. Some estimated results between HRNet and our method are shown in Fig.7. (a) In table.I, it shows the good generalization ability of our method. After adding our plug-in to each model, almost all performance have been improved. After adding PoseReNet, the average precision (AP) of SimpleBaseline-R50 has been increased by 0.5%, and the average recall rate (AR) increased by 0.6%. For R-101, its AP has increased by 0.7%, and its AR has increased by 0.2% after adding PoseReNet. For HRNet-w32 and input size 128×96, AP has been increased by 1.1% after adding PoseReNet, and for input size 256 × 192, AP has been increased by 0.6%.(b) After applied to the current best model, our method (table.II) achieved AP 77.3, 1.1% higher than the second performance. (c) From the table.III, our method improved the start-of-art model in accuracy in almost all key points on MPII. From the Fig.5, it can be found that the effect is well whether it is crowded or single-person scenes, or inconsistent character scales, even when the key points are occluded.

V. Conclusion

We present a light-weight but efficient Network to refine human pose estimation. We applied a channel attention mechanism to re-weight different scales of key points, and from extensive experiments, we can see that it can effectively improve the accuracy of the model. What's more, multi-scale feature map aggregation can provide more global information for training and inference stages. Our method serves as a light-weight model-agnostic plug-in, which can improve the accuracy with little computation cost. Used in mainstream models and data sets, significant improvements have taken place, indicating that our method has good generalization ability.

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[2] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[3] L. Xia, C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 20–27, 2012.

[4] J. Liu, H. Fu, and C.-L. Tai, "Posetween: Pose-driven tween animation," in *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, (New York, NY, USA), p. 791–804, Association for Computing Machinery, 2020.

[5] B. Wang, W. Liang, Y. Wang, and Y. Liang, "Head pose estimation with combined 2d sift and 3d hog features," in *2013 Seventh International Conference on Image and Graphics*, pp. 650–655, 2013.

[6] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[7] R. A. Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[8] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[9] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[10] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Computer Vision – ECCV 2016* (B. Leibe, J. Matas, N. Sebe, and M. Welling, eds.), (Cham), pp. 483–499, Springer International Publishing, 2016.

[11] Y. Cai, Z. Wang, Z. Luo, B. Yin, A. Du, H. Wang, X. Zhang, X. Zhou, E. Zhou, J. Sun, H. Bischof, T. Brox, and J.-M. Frahm, "Learning delicate local representations for multi-person pose estimation," in *Computer Vision – ECCV 2020*, (Cham), pp. 455–472, Springer International Publishing, 2020.

[12] G. Moon, J. Y. Chang, and K. M. Lee, "Posefix: Model-agnostic general human pose refinement network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[13] M. Ruggero Ronchi and P. Perona, "Benchmarking and error diagnosis in multi-instance pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[14] F. Zhang, X. Zhu, H. Dai, M. Ye, and C. Zhu, "Distribution-aware coordinate representation for human pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[15] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[16] W. Tang and Y. Wu, "Does learning specific features for related parts help human pose estimation?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[17] W. Tang, P. Yu, and Y. Wu, "Deeply learned compositional models for human pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[18] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[19] A. Newell, Z. Huang, and J. Deng, "Associative embedding: End-to-end learning for joint detection and grouping," in *Advances in Neural Information Processing Systems*, vol. 30, pp. 2277–2287, Curran Associates, Inc., 2017.

[20] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in

978-1-6654-4908-3/21/$31.00 ©2021 IEEE

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.

[21] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy, "Towards accurate multi-person pose estimation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4903–4911, 2017.

[22] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2334–2343, 2017.

[23] S. Huang, M. Gong, and D. Tao, "A coarse-fine network for keypoint localization," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3028–3037, 2017.

[24] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision – ECCV 2014*, (Cham), pp. 740–755, Springer International Publishing, 2014.

978-1-6654-4908-3/21/$31.00 ©2021 IEEE