Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

An Attention Module for Multi-Person Pose Estimation

Daxing Chen, Xinghao Song, Shixi Fan and Hongpeng Wang*
Harbin Institute of Technology,Shenzhen
Nanshan District, Shenzhen City, Guangdong Province, China
{*chendaxing,songxinghao*}*@stu.hit.edu.cn,*{*fanshixi,wanghp*}*@hit.edu.cn*

***Abstract*— In the top-down approaches of multi-person pose estimation, a human detector is adopted first to generate a set of human bounding boxes, then crop these human body and perform a single-person pose estimation model to get the final result. However, some body part of another person on the cropped image will interfere the single-person pose estimation model leading to an inaccuracy result. In order to model the relationship between adjacent keypoints effectively to alleviate this problem, we propose and attention module that could let the model get global receptive field at the shallow layer of the network and pay more attention to the key areas which is more important to pose estimation. Experiment results show that our method achieves 73.9% mAP with 2.4% absolute improvement compared to our baseline on the COCO test-dev dataset.**

***Index Terms*— Multi-Person Pose Estimation, Receptive Field, Attention Module.**

I. Introduction

Pose estimation is a basic research task of computer vision. It plays a fundamental role in other related fields of computer vision such as behavior recognition, character tracking and gait recognition. In real life, it can be applied to intelligent video surveillance, human-computer interaction, virtual reality, human animation, athlete-assisted training and so on. The problem of multi-person pose estimation has achieved great improvement with the development of deep convolutional neural networks [1], [2]. Multi-person pose estimation has two mainstream approaches. One is bottom-up framework [3]–[7] which all the human keypoints in the image are detected first, and then cluster them into different individuals. The other is top-down framework [8]–[14] which all the human bodies in the image are detected first by the human detection algorithm, and then crop the human bodies according to the bounding box to perform a single human pose estimation. The former is fast while the latter is highly accurate. In the top-down approaches, the cropped human body images may exist same body parts of other person and cause fail result. The relationship between adjacent keypoints can effectively alleviate this problem. In order to model the relationship between adjacent keypoints effectively,we proposes a module based on attention mechanism. This novel module which is combined with spatian attention,channel atention [15] and non-local operation [16], can enable the network to get global receptive field in the shallow layer, and pay more attention to the key areas related to the keypoints. Our works are as follows:

1) We propose a module combines attenton mechanism and non-local operation to effectively model relationship between keypoints to resolve the ambiguity caused by mutual human occlusion.
2) We explore the effects of the compoment in our proposed module involved in top-down framework.
3) Our proposed module makes a 2.4% absolute improvement compared to our baseline on the COCO test-dev dataset.

*Corresponding author

II. Related Work

The traditional human pose estimation models are basically based on the idea of template matching which need geometric prior. The core of these method is how to use the template to represent the whole human body structures,including the representation of keypoints, limb structures and the interrelationship between different limb structures. A good template could model more postures and get better result.

Deep convolutional neural networks have largely improved the performance of human pose estimation in recent years. In this paper, we mainly focus on the methods based on the convolutional neural network. The approaches of multi-person pose estimation could be divide into two categories: bottom-up approaches and top-down approaches.

A. Bottom-Up Approaches

The bottom-up approches detect all human keypoints first, and then cluster the detected keypoints into different individuals. Such approaches mainly focus on how to cluster keypoints to compose different individuals. DeepCut [3] proposes a partitioning and labeling formulation for a set of body-part hypotheses which generated with CNN-based part detectors. This formulation, an instance of an integer linear program(ILP), implicitly performs non-maximum suppression on the set of part candidates and groups them to form configurations of body parts respecting geometric and appearance constraints. It interprets the problem of distinguishing different persons in an image as an Interger Linear Program(ILP) problem and partitioning part detection candidates into person clusters. DeeperCut [4] improves DeepCut by using a strong body part detectors based on

978-1-7281-6321-5/19//$31.00 ©2019 IEEE

ResNet [1] and introducing a novel image-conditioned pairwise terms between body parts that push performance in the challenging case, and further boost the speed and robustness. Xia et al. [5] considered that human pose estimation and semantic part segmentation are two complementary tasks. They propose to solve the two tasks jointly for natural multi-person images, in which the estimated pose provides object-level shape prior to regularize part segments while the part-level segments constrain the variation of keypoints locations. Cao et al. [6] use part affinity fields(PAFs) to model the relationship between keypoints and assemble detected keypoints into different poses of people. PAFs is proposed to encode the limbs' position and orientation by a 2D vector, combined with the heatmap that marks the confidence of each keypoint, connect the keypoints of the same person by bipartite matching. The vector nature of the PAFs boost the performance. Kocabss et al. [7] propose a Pose Residual Network which introduces a single-shot object detection paradigm using grid-wise image feature maps. It represented body part proposals as region proposals, and detected limbs directly via a single-shot CNN. A bottom-up greedy parsing step is probabilistically redesigned to take global context into account.

B. Top-Down Approaches

Top-Down approaches locate and crop all persons from image first, and then perform a single-person pose estimation in the cropped person patches to get the result. Mask-RCNN [8] predicts human bounding boxes first and then crops out the feature map of the corresponding human bounding box to predict huamn keypoints. It models a keypoint's location as a one-hot mask, and predicts K masks,one for each of K keypoint types(e.g. left shoulder, right elbow). RMPE [9] uses Spatial Transformer Networks [10] handle inaccurate bounding boxes and redundant detections, extract high quality dominant human proposal, achieve further advances in performance.Cascade Pyramid Network(CPN) [11] mainly focus on problem that the difficulty of detecting different keypoint categories is different. It proposes a two-stage network contains GlobalNet and RefineNet to handle this problem. The GlobalNet learns a good feature representation based on feature pyramid network, it can provide sufficient context information,which is inevitable for the inference of the occluded and invisible joints. Based on these features, RefineNet detects the "hard" joints explicitly with online hard keypoints mining loss. Li et al. [12] propose to use multi-stage architecture to extract feature from image. With multi-level supervision and coarse-to-fine refinement, it achieves a high performance. Sun et al. [13]propose a High-Resolution Net(HRN), which is able to maintain high-resolution representations through the whole feature extraction process. It starts with a high-resolution subnetwork as the first stage, gradually adds high-to-low resolution subnetworks one by one to form more stages, and connects the multi-resolution subnetworks in parallel. High-resolution representations with multi-scale fusions help the method achieve state of the art. Simple Baseline [14] is a simple yet effective approach. It simply adds a few deconvolutional layers over the last convolution stage in the ResNet and gets comparable results. Because of it's simple and surprisingly effective, we adopt this model as our baseline.

C. Non-Local block

Convolutional operation could only process a local neighborhoods, either in space or time. Thus long-range dependencies are usually captured by stacking the operations repeatedly, propagating signals progressively through the network. But the network could only get the theoretical global receptive field in deep layer. Non-local [16] operation could help the network to get the global receptive field by computing the response at a position as a weighted sum of the feature at all positions in the input feature maps. According to [16], the non-local operation in deep neural networks is defined as:

$$y_i = \frac{1}{C(x)} \sum_{\forall_j} f\left(x_i, y_i\right) g\left(x_i\right). \tag{1}$$

where i is the index of the feature position whose response is to be computed and j is the index that enumerates all other positions. x is the input feature and y is the output feature with same size of x. f is a pairwise function which could calculate the relation between two points.The unary function g computes a representation of the input x at every position. The result is normalized by a factor $C\left(x\right)$. According to [16], the choices of f and g are not sensitive. In our work, we set g as a linear function which is implemented by 1x1 convolution, set f as Gaussian function $f\left(x_i, x_j\right) = e^{x_i^T x_j}$, set $C\left(X\right) = \sum_{\forall j} f\left(x_i, x_j\right)$. According to [16], a non-local block is defined as:

$$Z_i = W_z y_i + x_i \tag{2}$$

where y_i is given in (2) and "$+x_i$" denotes a residual connection.

D. Convolutional Block Attention Module

Visual attention mechanism has achieved great success in various tasks. A broad range of prior researches have investigated the spatial component and achieved great success. In recent years, the research on channel attention has received extensive attention. SE-Net [15] proposed a "Squeeze-and-Excitation" block to adaptively recalibrates channel-wise feature responses by explicitly modeling interdependencies between channels. Convolutional Block Attention Module(CBAM) [17] combined channel attention mechism and spatial attention mechism, proposed an effective attention module.Given and intermediate feature map $F \in R^{C \times H \times W}$

as input, CBAM sequentially infers a one dimention channel attention map $M_c \in R^{C\times1\times1}$ and a two dimention spatial attention map $M_s \in R^{1\times H\times W}$ to refine the feature. According to [17], the overall attenton process can be summarized as:

$$F^{'} = M_c(F) \otimes F, F^{''} = M_s\left(F^{'}\right) \otimes F^{'} \tag{3}$$

where $\otimes$ denotes element-wise multiplication. $F^{''}$ is the final refined feature. The channel attention map M_c is calculate by follow steps: first use both max-pooling and average-pooling to generating two different spatial context descriptors. Then forward these two descriptors to a shard network composed of multi-layer perceptron(MLP) with one hidden layer. After that, these two descriptors is composed by element-wise summation to get the final M_c. According to [17],M_c is computed as:

$$M_c(F) = \sigma\left(MLP\left(AvgPool\left(F\right)\right) + MLP\left(MaxPool\left(F\right)\right)\right) \tag{4}$$

The spatial attention module is calculated by follow steps: first apply max-pooling and average-pooling operation along the channel axis to generate two feature descriptors. Then concatenated these two feature descriptor across the channel. After that, feed them to a standard convolution layer to get the final spatial attention map. According to [17], M_s is computed as:

$$M_s(F) = \sigma\left(f^{7\times7}\left(\left[AvgPool\left(F\right); MaxPool\left(F\right)\right]\right)\right) \tag{5}$$

where σ denotes the sigmoid function and $f^{7\times7}$ represents a convolution operation with the filter size of 7×7.

III. Our Approach

In this paper, we focus on the top-down approaches to do multi-person pose estimation. Because same keypoint area often exits in the cropped human image's background, it's hard to distinguish the keypoint and the same keypoint of another people. So the network should have the ability to get large receptive field and capture long-range dependencies.The keypoint's detail information is also very important for the model to get the correct result. We combine non-local block and CBAM module to make the model have the ability to get global receptive field at shallow layer and focus on the features which would contribute to the result. We modify the CBAM module with global second-order pooling to further boost the attention module's capability.

CBAM module apply average-pooling and max-pooling to get feature descriptors. Neither average-pooling nor max-pooling operation just operate on a single channel. These operations ignore the relationship among channels.With the inspire of [18], which proposed a GSop block with global second-order pooling achieve a good performance in image recognition, we combine second-order pooling [27] with CBAM to capture the relationship between any two channels. Second-order pooling is an operation that implement by computing the outer product between a feature maps and its transpose, as Fig. 1 shows. With this operation, we could get an $c\times c$ convariance matrix. Each element of the convarizance encode the relationship of two channels. Compared to the max-pooling and the average-pooling, second-order pooling can capture the relationship of two channels better and provide more exact descriptors.We modify the CBAM module in the channel attention part. After modified,M_c is computed as:

$$M_c(F) = \sigma\left(MLP\left(F \times F^T\right)\right) \tag{6}$$

where σ denotes the sigmoid function and MLP is a multi-layer perceptron with two consecutive convolutions and non-linear activation to transform the convariance matrix to weight vector which is used to rescale the feature maps. $\times$ denotes outer product operation. We don't change the pooling operation in spatial attention because it would bring high computation cost and consume a lot of memory.

To enable the network's ability to get the global receptive field at the shallow layer, we combine non-local operation to our attention module. As a result,our module can be describe as:

$$Refined(F) = Z\left(M_s\left(M_c(F) \otimes F\right) \otimes \left(M_c(F) \otimes F\right)\right) \tag{7}$$

where Z is given in (2), M_s is given in (5), M_c is given in (6). F is the input feature maps. $Refined(F)$ denotes the refined future map.

IV. Experiments

All of our experiments are base on the Simple Baseline model implemented under the Pytorch framework, and runs on one workstations which is equipped with two NVIDIA Titan V GPUs and one Intel i7-7800X@3.50GHz CPU. We do our experiments on COCO [19] dataset. COCO dataset contains over 200,000 images and 250,000 person instances labeled with 17 keypoints. Our models are only trained on COCO train2017 dataset which includes 57K images and 150K person instances. We validate our models on COCO val2017 dataset which includes 5000 images and 6352 person instances. We test our model on COCO test-dev dataset which include 20K images and submit the result to the evaluate server to get the final scores. We use official evaluation metric of COCO keypoit detection task which is average precision based on object keypoints similarity(OKS) [20]. The OKS defines the similarity between the predicted pose and the ground truth pose which play the same role as th IoU in object detection task.

In our experiments, We embed our module into ResNet after the $C2$, $C3$, $C4$, $C5$ layers. It is noted that the module can be inserted to anywhere at neural network. But if the module is embed at very shallow layer, it would cause large memory consume and computation cost.

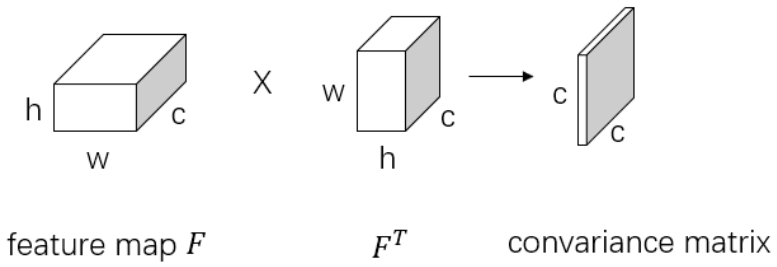

Fig. 1. The second-order pooling operation

A. Training and testing strategy

Follow the prior works, we extend the aspect ration of human bbox to 4:3 and then crop the human according to the bbox from the image. The cropped images are resized to a fixed size,256x192 or 384x288. Random ratation$(-40°, 40°)$, random scale$[0.7, 1.3]$) and flipping are used to do data augment. We use the Adam [21] as optimizer. The init learing rate is setted as 1e-3 and dropped to 1e-4 and 1e-5 at the 90th and 120 epochs respectively. The ResNet's layers are initialized with the weight pretrained on ImageNet [22] while our module is trained from scratch.The model is trained 140 epochs. We use the same person detect result provided by Simple Baseline [14] for both validataion set and test-dev set.

B. Results on the validation set

We report the results of our method in Table I. It is noted that the parameters and GFLOPs are only calculated for the pose estimation network. Our method almost train from scratch with the input size 256x192 achieves an 72% AP score. Comparing to the corresponding baseline with same input size, out method improves AP by 1.6% while the increment of computation cost is little. The increment of parameters is also acceptable. Our method achieves the same performance of baseline with the backbone ResNet-152 which has about 1.6 times parameters and 1.58 times GFLOPs compare to ours.

While the input size is 384x288, our method gets a 74.5% AP, which have 2.3% improvements comparing with corresponding baseline. This result even outperform the baseline with ResNet-152 backbone by 0.2% while the parameters and GFLOPs is less. We should note that our model is totally train from scratch while the baseline's backbone is pretrain on ImageNet. We believe that our method's accuracy could be further boost if it's pretrained on ImageNet.

C. Results on the test-dev set

Table II reports the experiment result on test-dev dataset. Our method with input size 256x192 achieves an AP of 71.6%, which have 1.6% improvement. Our method with input size 384x288 achieves an AP of 73.9% with a 2.4% improvement and even outputperform the baseline with a stronger backbone.

Table III reports the performances of our approaches and the existing state-of-the-art approaches. The data of other approachs is referenced from [13]. Our approaches outperform all the bottom-up approaches and most of the prior top-down approaches. Without pretrained on ImageNet, our approach achieves a 73.9% AP, which is only 1.6% lower than HRNet-w48 while it's parameters is 21.3M more than ours and calculated cost is 10.6 GFLOPs more than ours.

D. Ablation study

Table IV reports the effect of each component in our module. All results are obtained over the input size of 256x192 with backbone ResNet-50 on the COCO validation dataset. All added modules is embedded into ResNet after $C2$, $C3$, $C4$, $C5$ layers. It can be seen that while we embeded the original CBAM module into ResNet, the AP drop to 70.2%, which is 0.2% lower than baseline. It shows that the original CBAM module couldn't bring benefit to the multi-person pose estimation task.While we embeeding non-local block into ResNet, the module achieves a 71% AP, which get 0.6% improvement. It shows that the global receptive field is important to multi-person pose estimation task. The CBAM module with modified by second-order pooling make a great improvement to the model which achieves a 71.8% AP. Compared to the origin CBAM, second-order pooling improve it's capability by capturing the relationship of two channels more exact. The combination of non-local and CBAM block with second-order pooling further boost the model's performance to 72% AP. It's obvious that our proposed module is efficient yet effective.

TABLE I

COMPARISONS ON THE COCO VALIDATION SET. RESNET-50* DENOTES THE RESNET WITH OUR PROPOSED MODULE

Method	Backbone	input size	Params	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
Baseline	ResNet-50	256x192	34M	8.9	70.4	88.6	78.3	67.1	77.2	76.3
Baseline	ResNet-101	256x192	53M	12.4	71.4	89.3	79.3	68.1	78.1	77.1
Baseline	ResNet-152	256x192	68.6M	15.7	72	89.3	79.8	68.7	78.9	77.8
ours	ResNet-50*	256x192	42.3M	9.9	72	89.4	79.6	68.8	78.8	77.9
Baseline	ResNet-50	384x288	34M	20.2	72.2	89.3	78.9	68.1	79.7	77.6
Baseline	ResNet-152	384x288	68.6M	35.6	74.3	89.6	81.1	70.5	79.7	79.7
ours	ResNet-50*	384x288	42.3M	22.3	74.5	90	81.2	70.6	81.8	79.8

TABLE II

COMPARISONS ON THE COCO TEST-DEV SET.

method	backbone	input size	params	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
Baseline	ResNet-50	256x192	34M	8.9	70	90.9	77.9	66.8	75.8	75.6
ours	ResNet-50*	256x192	42.3M	9.9	71.6	91.5	80	68.7	77.2	77.2
Baseline	ResNet-50	384x288	34M	20.2	71.5	91.1	78.7	67.8	78	76.9
Baseline	ResNet-152	384x288	68.6M	35.6	73.7	91.9	81.1	70.3	80	79
ours	ResNet-50*	384x288	42.3M	22.3	73.9	91.9	81.9	70.6	80	79.2

TABLE III

COMPARISONS WITH OTHER METHOD ON THE COCO TEST-DEV SET.

method	backbone	input size	params	GFLOPs	AP
Bottom-up approaches					
OpenPose [6]	-	-	-	-	61.8
Associative Embedding [23]	-	-	-	-	65.5
PersonLab [24]	-	-	-	-	68.7
MultiPoseNet [7]	-	-	-	-	69.6
Top-down approaches					
Mask-RCNN [8]	ResNet-50-FPN	-	-	-	63.1
G-RMI [25]	ResNet-101	353x257	42.6M	57	64.9
CPN [11]	ResNet-Inception	384x288	-	-	72.1
RMPE [9]	PyraNet	320x256	21.8M	26.7	72.3
CFN [26]	-	-	-	-	72.6
Baseline	ResNet-152	384x288	68.6M	35.6	73.7
HRNet-w48 [13]	HRNet-w32	384x288	63.6M	32.9	75.5
ours	ResNet-50*	384x288	42.3M	22.3	73.9

TABLE IV

ABLATION STUDY RESULT.

Method	Params	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
Baseline	34M	8.9	70.4	88.6	78.3	67.1	77.2	76.3
+ CBAM	35.4M	8.99	70.2	88.7	77.9	67.1	76.8	76.3
+ NL	39.6M	9.74	71	88.7	78.5	67.8	77.5	76.7
+ CBAM with 2^{nd} order pooling	36.7M	9.17	71.8	89.2	79.6	68.5	78.6	77.7
ours	42.3M	9.9	72	89.4	79.6	68.8	78.8	77.9

V. Conclusion

In this paper, we propose an attention module to combine spatial attention mechanism and channel attention mechanism which is improved by global second-order pooling. In addition, we introduce a non-local module in the attention module to make the network have ability to get global receptive field in shallow layer. The experimental results show that the proposed module can improve the accuracy of the model effectively on the COCO dataset.

References

[1] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[2] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

[3] Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P. V., & Schiele, B. (2016). Deepcut: Joint subset partition and labeling for multi person pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4929-4937).

[4] Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., & Schiele, B. (2016, October). Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In European Conference on Computer Vision (pp. 34-50). Springer, Cham.

[5] Xia, F., Wang, P., Chen, X., & Yuille, A. L. (2017). Joint multi-person pose estimation and semantic part segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6769-6778).

[6] Cao, Z., Simon, T., Wei, S. E.,& Sheikh, Y. (2017). Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7291-7299).

[7] Kocabas, M., Karagoz, S., & Akbas, E. (2018). Multiposenet: Fast multi-person pose estimation using pose residual network. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 417-433).

[8] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).

[9] Fang, H. S., Xie, S., Tai, Y. W., & Lu, C. (2017). Rmpe: Regional multi-person pose estimation. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2334-2343).

[10] Jaderberg, M., Simonyan, K., & Zisserman, A. (2015). Spatial transformer networks. In Advances in neural information processing systems (pp. 2017-2025).

[11] Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., & Sun, J. (2018). Cascaded pyramid network for multi-person pose estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7103-7112).

[12] Li, W., Wang, Z., Yin, B., Peng, Q., Du, Y., Xiao, T.& Sun, J. (2019). Rethinking on Multi-Stage Networks for Human Pose Estimation. arXiv preprint arXiv:1901.00148.

[13] Sun, Ke, et al. "Deep high-resolution representation learning for human pose estimation." arXiv preprint arXiv:1902.09212 (2019).

[14] Xiao, B., Wu, H., & Wei, Y. (2018). Simple baselines for human pose estimation and tracking. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 466-481).

[15] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7132-7141).

[16] Wang, X., Girshick, R., Gupta, A., & He, K. (2018). Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7794-7803).

[17] Woo, S., Park, J., Lee, J. Y., & So Kweon, I. (2018). Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 3-19).

[18] Gao, Z., Xie, J., Wang, Q., & Li, P. (2019). Global Second-order Pooling Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3024-3033).

[19] Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D, & Zitnick, C. L. (2014, September). Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Springer, Cham.

[20] COCO keypoint evaluation website.`http://cocodataset.org/#keypoints-eval`

[21] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

[22] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.

[23] Newell, A., Huang, Z., & Deng, J. (2017). Associative embedding: End-to-end learning for joint detection and grouping. In Advances in Neural Information Processing Systems (pp. 2277-2287).

[24] Papandreou, G., Zhu, T., Chen, L. C., Gidaris, S., Tompson, J., & Murphy, K. (2018). Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 269-286).

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

[26] Huang, S., Gong, M., & Tao, D. (2017). A coarse-fine network for keypoint localization. In Proceedings of the IEEE International Conference on Computer Vision (pp. 3028-3037).

[27] Carreira, J., Caseiro, R., Batista, J., & Sminchisescu, C. (2012, October). Semantic segmentation with second-order pooling. In European Conference on Computer Vision (pp. 430-443). Springer, Berlin, Heidelberg.