DoubleHigherNet: Coarse-to-Fine Precise Heatmap Bottom-Up Dynamic Pose Computer Intelligent Estimation

Yiheng Peng[1, *], Zhichun Jiang[2]

[1]School of Computer Science and Technology, Donghua University, Shanghai, China, 201620
[2]Department of Statistics, University of Strathclyde, Glasgow, England

*Corresponding author: yihengpeng@dhu.edu.cn

Abstract. Accurate keypoint positioning is necessary for bottom-up multi-person pose estimation methods to handle scale variation and crowdedness. In this paper, we present DoubleHigherNet: a novel network learning scale-aware and precise heatmap representation for bottom-up process using double high-resolution feature pyramids and coarse-to-fine training. The two feature pyramids in DoubleHigherNet consists of 1/ 4 resolution feature and higher-resolution (1/2) maps generated by attention fusion blocks and transposed convolutions. Benefited by the training strategy, muti-resoltion and coarse-fine heatmap aggregation, the proposed approach is able to predict keypoints more accurately so as to perform better on difficult crowded scenes. DoubleHigherNet-w32 achieves competitive result on CrowdPose-test, surpassing all the top-down methods and bottom-up SOTA HigherHRNet-w32 (which possesses similar number of params with DoubleHigherNet-w32).

Keywords: DoubleHigherNet, Coarse-to-Fine training, Heatmap aggregation, Attention fusion block.

1. Introduction

2d multi-person pose estimation aims to locate the pixel coordinates of human body parts (elbows, wrists, etc.) on 2d images, and use them as the basis for many computer vision tasks, such as: action recognition, human-computer interaction, and person Re-ID. Benefited by deep learning and informative benchmarks such as COCO and CrowdPose [4], great advance has been witnessed in this field.

The human pose estimation family consists of two mainstream factions：top-down and bottom-up. A top-down method first employs a human detector such as Mask-Rcnn (He et al.2017) to obtain the bounding-box of each person instance in the image. Every person bounding-box is then cut out from the original image and be input into a single-person key-point detection network. Since top-down methods normalize every person instance to similar scale by resizing the cropped bbox, they are more robust to scale variations. However, due to top-down methods' relying on person bounding-box, they have intrinsic disadvantage for dense scenes where bounding-boxes tend to overlap. What's more, such methods have to estimate joints (keypoint) positions for every human separately, the computational cost

of the algorithm will increase greatly as the number of people in the photo increases. In contrast, the bottom-up process first determines the identity-free joints position of all people in the input image by predicting the heatmaps of different body parts, and then groups them into instances of different people. This strategy effectively improves the speed of bottom-up methods and their ability to realize real-time pose estimation. Independent of human detector, bottom-up methods perform better on crowd-pose, a benchmark with various dense and difficult scenes. Compared with top-down, the disadvantage of bottom-up process is that it estimates the key points of all people in a photo simultaneously, which will be influenced by the change of the scale between people.

In order to tackle the problem of inaccurate body parts positioning caused by scale variations, two strategies have been proved to be effective in previous bottom-up models. PersonLab increase the size of heatmap for grouping by utilizing high-resolution input images. As there is a conflict between estimating small persons and large persons, the second strategy, feature pyramid, is introduced by HigherHRNet [3] to balance the performance on persons of different scales.

In this paper, we introduce coarse-to-fine heatmap training strategy to bottom-up process, while remaining the first two strategies as well. As far as we know, in the field of bottom-up, we are the first to take all the three strategies into consideration. We propose a novel network, DoubleHigherNet, for both heatmap and BU grouping component regression. DoubleHigherNet contains two high-resolution feature pyramid modules. Unlike some feature pyramid structures, in our model, the size of Gaussian kernel will not vary with the scale of different feature maps in the same feature pyramid. The first pyramid structure aims at learning general positioning information and regressing coarse key-points heatmaps with larger Gaussian kernel sizes (coarse) while the second feature pyramid is required to locate key points more accurately (fine). The two feature pyramids are connected by several residual blocks which further study precise positioning information. High-resolution feature maps are derived by attention fusion block and deconvolution. The proposed fusion block either fuses rough and fine locating information or fuses semantic information and 1/4 resolution heatmaps. Making full use of all the heatmaps we generate, we improve HigherHRNet [3]'s heatmap aggregation strategy to obtain accurate heatmaps.

We verify the effectiveness of our model on the challenging CrowdPose [4] dataset and the model achieves competitive results among models with similar amount of parameters. Ablation experiments prove the effectiveness of coarse-to-fine learning in the bottom-up method. Moreover, we find that if coarse and fine heatmaps are aggregated, the estimation will become more accurate.

The main contribution of this paper is concluded as follows:

We attempt to generate precise key-points heatmap representation for bottom-up process.

We propose a DoubleHigherNet with two cascaded feature pyramids. We introduce coarse-to-fine training strategy for it to improve the accuracy and robustness of heatmaps it generates. We find that the aggregation of coarse and fine heatmaps is helpful for key points positioning.

We propose a fusion block that plays a function analogous to but better than concat operation.

Our model achieves competitive results among models with similar amount of parameters on the challenging CrowdPose [4] dataset with a variety of testing scenes such as occlusion, crowdedness and body part deformation.

2. Related work

2.1. Top-down methods

Top-down methods detect the joints positions of a single person instance in the bbox generated by a person detector. Mask R-CNN (He et al.2017) adds an extra key point detection sub-branch to Faster R-CNN, and reuses features after ROI-Pooling. CPN [7] proposes a refine-net to further study the heatmap learned by global-net. Stacking the global-net part of CPN, MSPN [6] integrates coarse-to-fine training and intermediate supervision. It employs cross-level information fusion to achieve better results. Online hard key-points mining is adopted by both CPN and MSPN. Qiu et al. propose OpecNet [5] based on Gcn to learn the graph information of human body.

2.2. Bottom-up methods

Bottom-up methods detect the identity-free body part key points of all people in the image and groups them into full pose. Associate-Embedding [2] uses muti-stage stacked-hourglass network for both heatmap prediction and grouping components regression. The grouping process is realized with the help of associate-embedding, which assigns a "label" (tag) to each key point. PersonLab uses extended ResNet and groups body joints by directly learning and regressing a 2D displacements field for each key points pair. HigherHRNet [3] adds a two-level feature pyramid structure after HRNet [1] to gain scale-aware high-resolution heatmaps, and its grouping process is the same as that of AE

2.3. Coarse-to-fine training

Coarse-to-fine training was originally used in top-down methods. In global-net, CPN [7] uses different Gaussian kernel size across feature maps. Finally, the kernel size is reduced to 7 in refine-net to learn precise key points positioning. With the observation of heatmaps' becoming more accurate stage by stage, MSPN [6] uses bigger kernel size in former stages and smaller kernel in latter stages to cater to the network's heatmap positioning characteristics and to perform intermediate supervision.

Figure 1: the left three heatmaps are used as coarse ground truth and the right ones are for fine training.

Our work is partly inspired by MSPN [6]. However, we simply adopt coarse-to-fine training for the sake of complying with the characteristics of the network and obtaining accurate heatmaps. Intermediate supervision is not necessary for HRNet [1] based networks.

2.4. Keypoints regression head

Part of the bottom-up networks are obtained by adding an extra Keypoints regression head after a single-person pose detector network (HRNet, Hourglass, etc.). The most basic regression head is a simple 1×1 convolutional layer. Such strategy is adopted by Hourglass+ Associate-Embedding. The regression head of bottom-up HRNet is slightly more complicated. Features of four scales output by HRNet [1] are upsampled to 1/4 and are concatenated together before applying a 1×1 convolution. HigherHRNet [3] only retains the highest resolution feature maps output by HRNet. It uses a 1×1 convolution to regress heatmap at resolution 1/4 and then performs cat fusion between the heatmap and the 1/4 resolution feature. After the concatenation operation, fused information is input into the deconvolution module, and heatmap at resolution 1/2 of the input image is then obtained by another 1×1 convolution.

HRNet [1] is also used as backbone in this paper. For the sake of coarse-to-fine training, the keypoint-head behind the backbone is a double-stage cascaded feature pyramid structure. As to further study the accurate position of joints, the two stage are connected by several Bottlenecks. We also use deconvolution to obtain a high-pixel heatmap.

3. Double-Higher Network

3.1. DoubleHigherNet

DoubleHigherNet contains two high-resolution feature pyramid modules. The two feature pyramids are built based on the 1/4 resolution path of HRNet [1]**.** Thus, except for regression results and fused features output by fusion blocks, all the feature in the network share the same channel num of *C*, set as 32 or 48 in HRNet. Following HigherHRNet [3], deconvolution layers are used for generating feature maps 2 times larger in resolution than the input. With the help of fusion block, 4 × 4 deconvolution and 1 ×

1 convolution, coarse heatmaps of two level (1/4 and 1/2) are predicted by stage 1. In stage 1, fusion block fuses high-level semantic information and 1/4 resolution heatmaps as input for deconvolution.

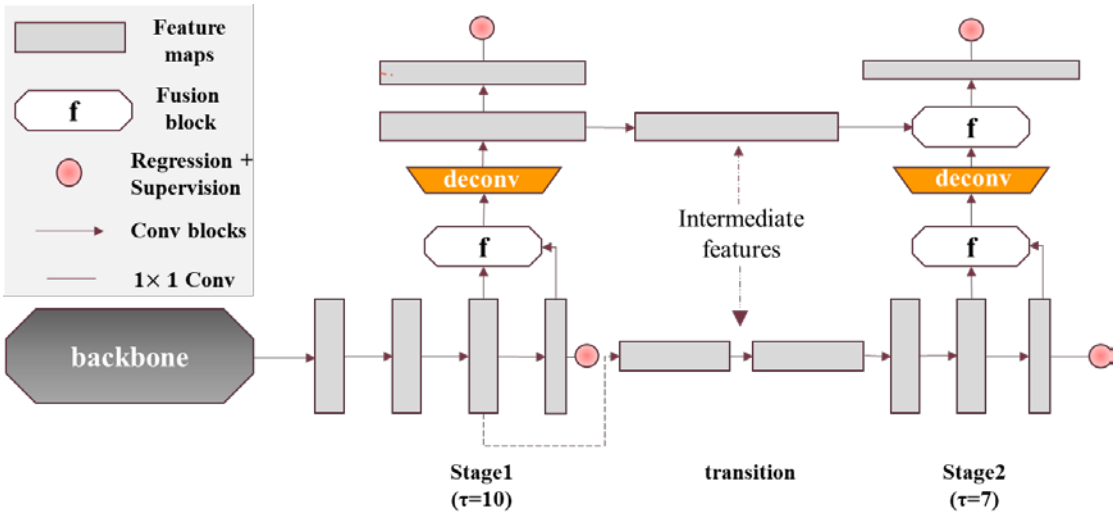

Figure 2: An illustration of DoubleHigherNet

Thus, a reduction of channel after deconvolution can be realized and the channel num remains to be relatively high and stable, which benefits the performance of upsampling.

The transition stage contains different numbers of BottleNeck modules at the two levels. It functions as further studying precise heatmap positioning information from the coarse features. Different from fusion blocks before deconv, the last fusion block fuses both rough and accurate locating information for precise heatmap regression. Finally, fine heatmaps at 1/4 and 1/2 resolution are predicted in stage 2.

3.2. Coarse-Fine Heatmap

It is worth noting that directly regressing the coordinates of human body parts is unreasonable and too intuitive. Thus, in an analogous thoughts of label smoothing, heatmap representation is introduced as network's output target. Assigning non-zero confidence value to pixels close to a keypoint, heatmap encodes spatial contextual clues and takes possible target ambiguity into consideration. Heatmaps are generated by modeling the actual kepoint positions as Gaussian peaks. Let $hm_j \in R^{Width \times Height}$ $(k = 1,2, \dots, J)$ denote the heatmap of the j-th kind of joint and J is the number of joint categories annotated in a certain dataset. For a position (x, y) in the input image, hm_j is calculated by the formula we list below:

$$hm_j(x,y) = \begin{cases} exp(-\dfrac{||(x,y) - (x_j^i, y_j^i)||_2^2}{\sigma^2}), & (x,y) \in \aleph_j^i \\ 0, & otherwise \end{cases}$$

In which σ is a constant controlling the variance of Gaussian distribution. Ground-truth heatmaps from different stages and levels are applied with the same σ, set as 7 in our implementation. $\aleph_j^i = (x,y)|\ ||(x,y) - (x_j^i, y_j^i)||_2 \leq \tau$ defines the regression scope influenced by Gaussian kernel. Using a large τ means allowing the network to locate joints relatively inaccurately and vice versa. Experimental results show that heatmaps predicted by CNN tend to become precise layer by layer. In order to cater to this characteristic, we set τ as 10 for 1/4 and 1/2 resolution heatmaps in stage1. For all the ground-truth heatmaps in stage2, τ is set as 7, which is relatively smaller. Intuitively, the ground-truth heatmaps of

stage1 will be coarser and the ones of stage2 will be more accurate. Such setting makes it easier for the network to refine the predicted heatmaps gradually. MSE loss is adopted while heatmap training.

3.3. Attention-Fusion Block

We propose a channel-aware fusion block which assigns reasonable weights to different feature channels while concatenating. This fusing method is partly inspired by OPECNet [5]. In a fusion block, we first concatenate the two feature map to be fused and then input the concatenated feature into a channel-attention layer (a.k.a SElayer) for weights obtaining. A channel-attention layer usually consists of global average pooling, two fully connected layers and a sigmoid layer. Global average pooling serves as reducing the width and the height of the map into 1. Then, the network is able to focus on the dimension of channel and the weight is the same for every pixel in a certain channel. To reduce parameters and increase nonlinearity, reduction operation is adopted by the first Fc layer. The reduction ratio is set as 4 in this paper. The second Fc layer recovers the channel num to be consistent with the original concatenated feature. After that, a sigmoid layer maps the features into weights between $(0, 1)$.

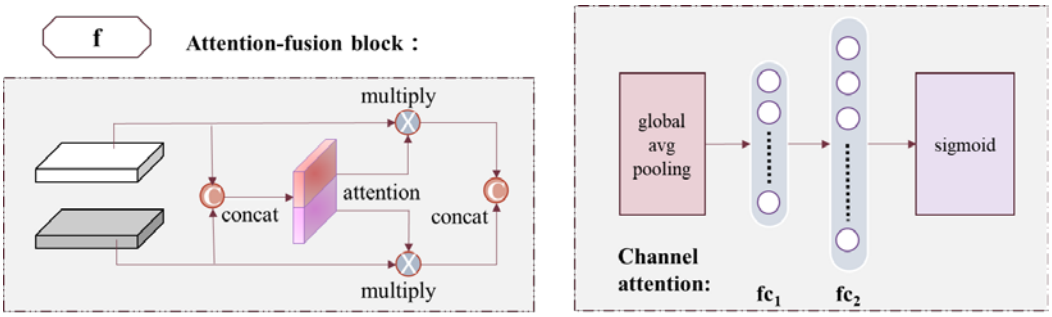

Figure 3: An illustration of attention fusion block

Finally, we multiply the weights with the corresponding channels from the two feature maps and perform cat operation. There is no hyper-parameters needed in the fusion block and weights learning is also free from extra supervision with the help of channel attention. We use three fusion blocks in DoubleHigherNet, two for fusing feature and heatmaps at 1/4 resolution, and one for integrating rough and fine locating information before predicting the last heatmap.

3.4. Tag-map

We use associate-embedding as grouping method for a fair comparison with HigherHRNet [3]. AE is actually a feature map (tagmap) that corresponds to a certain heatmap at pixel level. Pixels belonging to the same person have similar tags while tags of pixels in different person instances have significant difference. Following HigherHRNet, we only regress tagmaps in the first level, i.e., at 1/4 resolution of input image. Only the tagmap predicted by stage2 is used for grouping.

3.5. Heatmap Aggregation

It is pointed out by HigherHRNet [3] that averaging heatmaps from different level(scale) can complement their ability to locate joints of different scales. This strategy is adopted in our work to generate scale-aware heatmap. We find that aggregating coarse and fine heatmaps also benefits joints positioning. First of all, it's easier for the coarse heatmap to detect a hard keypoint, in another word, to assign greater confidence value around the joint position. Secondly, despite the smaller confidence value, fine heatmap predict accurate joint position. Thus, making full use of both kinds of heatmaps, such aggregation strategy enables the pixels around a joint to have large response value (compared with pure fine heatmap) without harming precise keypoint positioning.

We perform the coarse-fine heatmap aggregation before using bilinear interpolation. As a result,

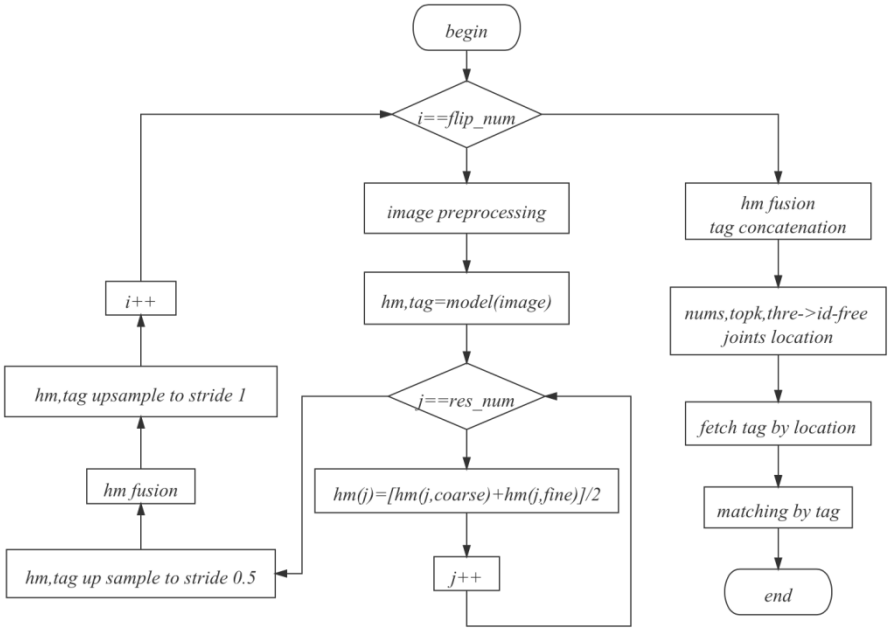

Figure 4: Flow chart for inference process

Only one 1/4 resolution heatmap is required to be upsampled to 1/2. After that, heatmaps from the two scale are averaged at resolution of 1/2 of the input image. Heatmaps are then further upsampled to the original image size for flip test averaging. Tagmaps are also upsampled but are not averaged. Due to the flip test, two 1-d tags (label) are concatenated into a 2-d tag so that the l_2 distance between tags can be calculated. According to the joints locations obtained by NMS, we fetch the corresponding tags from the tagmap and group kepoints whose tags have small Euclidean distance together.

4. Experiments

4.1. CrowdPose

CrowdPose [4] is a challenging bench mark with a variety of testing scenes such as occlusion, crowdedness and body part deformation. It contains 20,000 images and about 80,000 person instances. The training, validation and testing sub-dataset are split in proportional to 5:1:4. The standard evaluation metric for CrowdPose is based on Object Keypoints Similarity (OKS):

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

In which d_i is the l_2 distance between the predicted joints and ground truth, v_i is the visibility flag of the ground-truth keypoint, s is the sacle of the instance, and j_i is a per-keypoint constant that controls falloff. The standard average precision and recall scores are shown as follow: AP^{50}(AP at OKS = 0.50), AP^{75}, mAP (the mean of AP scores at 10 positions, OKS = 0.50, 0.55, …, 0.90, 0.95), AP^{easy} $(crowd\ index \in (0{,}0.1))$, $AP^{medium}(0.1 - 0.8)$, $AP^{hard}(0.8 - 1)$ and mAR at OKS = 0.50, 0.55, …, 0.90, 0.95.

4.2. Implementation

We follow the conventional data augmentation for CrowdPose [4] dataset. We adopt random rotation of ($[-30^{\circ}, 30^{\circ}]$), random scaling of $[0.75, 1.5]$, random translation of $[-40, 40]$, and random horizontal flip

at a probability of 0.5. All the images are resized to 512×512 during training. We use both the training-set and the validation-set for network training and conduct ablation study and model comparison on crowdpose-test. We take HRNet-w32 as backbone and use Adam optimizer for parameters learning. The initial learning rate is set as 1e-3 and decrease by multiplying 0.1 at the 200th and 260th epoch. Following HigherHRNet [3], we train the model for a total of 300 epochs and adopt flip test for all the experiments.

4.3. Results on CrowdPose

Table 1: Comparison with other state-of-the-arts on CrowdPose-test

Arch	mAP	AP^{50}	AP^{75}	AP^{E}	AP^{M}	AP^{H}
Openpose	-	-	-	62.7	48.7	32.3
Mask-RCNN	57.2	83.5	60.3	69.4	57.9	45.8
RMPE	61.0	81.3	66.0	71.2	61.4	51.1
SPPE	66.0	84.2	71.5	75.5	66.3	57.4
HigherHRNet-w32	66.1	86.4	71.0	74.2	67.0	56.6
DoubleHrNet-w32	66.8	87.1	71.7	75.0	67.7	57.5

In table 1, we compare DoubleHigherNet with other state-of-the-arts on CrowdPose-test. Our method achieves competitive results without using refinement or other post-processing methods. DoubleHigherNet outperforms HigherHRNet-w32 [3], which possesses similar number of params with DoubleHigherNet-w32, by 0.9% AP for hard scenes (AP^{H}). This observation verifies the model's ability to predict accurate heatmaps and robustness to challenging scenes. Compared to SPPE [4], our model has a stronger coarse positioning capability with an improvement of AP^{50} by 2.9%.

Table 2: Ablation study on CrowdPose-test

$fusion\ block$	$coarse2fine$	$Aggregation$	mAp	AP^{50}	AP^{75}	mAR
			64.0	84.9	68.8	70.4
√			65.1	85.9	69.9	71.0
√	√		65.3	85.9	70.0	71.3
√	√	√	65.5	86.1	70.2	71.4

We conduct ablation study on CrowdPose-test as the validation set is already used as training. We evaluate the impact of our proposed attention fusion block, coarse-to-fine learning and coarse-fine heatmap aggregation. Results are shown in table 2. All the heatmap fusing strategies will fail without fusion blocks. Thus, by introducing attention fusion block, the performance improves to 65.1% with 1.1% mAP increasing. Coarse-to-fine training improves 0.2% mAP (τ = 7 only). Based on the coarse and fine heatmaps, our aggregation strategy further improves 0.3% mAP. The result shows the effectiveness of our fusion block, coarse-to-fine training and our coarse-fine heatmap aggregation strategy.

4.4. Qualitative results

Qualitative results on CrowdPose [4] are shown in figure 5. We draw the human skeletons of multiple people extracted by the algorithm. The qualitative results further verifies the effectiveness of our model in complex scenes, e.g., body part deformation and cluttered background (1st example), small-scale pose estimation (2nd example), person-overlapping (3st and 4st examples), scale-variation (7st example), and crowded scene (3rd , 5th and 8th examples).

Figure 5: Qualitative results on CrowdPose

5. Conclusion

In this paper, we present DoubleHigherNet: a novel network designed for bottom-up muti-person pose estimation. It learns scale-aware and precise heatmap representation using double high-resolution feature pyramids and coarse-to-fine training. The two feature pyramids in DoubleHigherNet consists of 1/4 resolution feature and 1/2 resolution maps generated by attention fusion blocks and transposed convolutions. We find that the training strategy, muti-resoltion and coarse-fine heatmap aggregation enable the model to predict keypoints more accurately and to perform better on difficult crowded scenes. DoubleHigherNet-w32 surpasses all the top-down methods and bottom-up SOTA HigherHRNet-w32 on CrowdPose, especially for those challenging scenes.

Reference

[1] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, MingkuiTan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recogntion. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

[2] Alejandro Newell, Zhiao Huang, and Jia Deng. Associa-tive embedding: End-to-end learning for joint detection and grouping. In NeurIPS, 2017.

[3] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi,Thomas S. Huang, and Lei Zhang. Higherhrnet: Scale-awarerepresentation learning for bottom-up human pose estimation. In CVPR, 2020.

[4] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In CVPR, 2019.

[5] Lingteng Qiu, Xuanye Zhang, Yanran Li, Guanbin Li, Xiaojun Wu, Zixiang Xiong, Xiaoguang Han, and Shuguang Cui: Peeking into occluded joints: A novel framework for crowdpose estimation in ECCV, 2020.

[6] Wenbo Li, Zhicheng Wang, Binyi Yin, Qixiang Peng, Yuming Du, Tianzi Xiao, Gang Yu, Hongtao Lu, Yichen Wei, Jian Sun: Rethinking on Multi-Stage Networks for Human Pose Estimation. arXiv preprint, 2019

[7] Yilun Chen, Zhicheng Wang, Y uxiang Peng, Zhiqiang Zhang, Gang Y u, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In CVPR, 2018