ACPNET:ANCHOR-CENTER BASED PERSON NETWORK FOR HUMAN POSE ESTIMATION AND INSTANCE SEGMENTATION

Yang Bai [1], *Weiqiang Wang*[1]

[1] University of Chinese Academy of Sciences,CAS, Beijing, China
Email:baiyang17@mails.ucas.edu.cn, wqwang@ucas.ac.cn

ABSTRACT

We present an effective approach to tackle the multi-person pose estimation and person instance segmentation jointly. The approach based on Mask R-CNN uses a set of well-designed labels, called anchor-center based label, to learn keypoints localization in complex and crowded multi-person scenes. Combining the annotation of bounding boxes, we adopt a regression method to predict heatmaps and 2D-offset vector fields based on anchor-center for each keypoint type. Instead of using a person-detector, we use the person proposal network to predict anchors' locations in images. To generate high-quality segmentation mask, we use the ResNet-FPN with the deformable convolutions to model geometric transformations for non-rigid objects. Our method can efficiently and effectively deal with human pose estimation and instance segmentation tasks in a flexible end-to-end manner. Without bells and whistles, our method achieves a comparable result on the COCO keypoints task and the state-of-the-art accuracy on the COCO person instance segmentation task.

Index Terms— Person Proposal Network, Anchor-Center based Label, Deformable, Regression, Multi-Task

1. INTRODUCTION

Recently, visual understanding, such as object detection, semantic segmentation and human pose estimation, has witnessed a significant progress due to deep learning. Especially, understanding human activities in images or videos plays a basic role in a variety of visual tasks. Multi-person pose estimation and person instance segmentation in unconstrained environments present a unique set of challenges. First, it is difficult to accurately locate keypoints and make association due to limb articulations, occlusion, crowd, especially in complex multi-person scenes. Second, The computation cost of a system increases quickly with the number of people in images, which makes making practical AI products a challenge. Third, by far no framework can implement pose estimation and instance segmentation efficiently and effectively in an end-to-end manner. The goal of this work is to develop a comparably efficient and effective framework for locating the keypoints associated with each person and estimating its instance segmentation mask jointly.

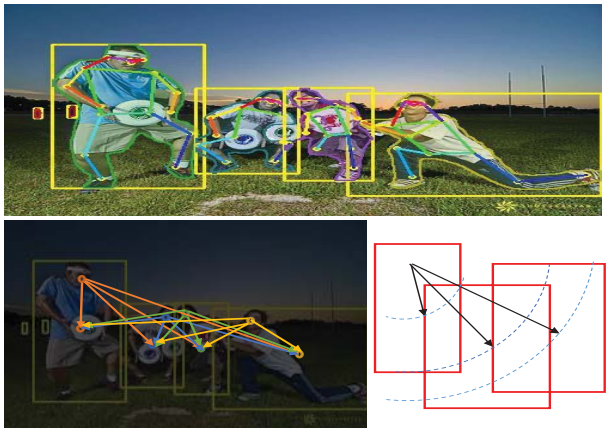

Fig. 1. **Top**: Multi-person pose estimation and person instance segmentation. **Bottom left**: Anchor-center based labels and 2D-offset vector fields. **Bottom Right**: A simple geometric diagram shows the predicted label and offset vector fields (keypoint:the leftmost man's head in **Top** image).

The bottom-up approaches [1, 2, 3, 4, 5] detect all keypoints and use the sophisticated assignment algorithm to group keypoints together into specific instance. They decouple the time complexity from the number of people in images. Yet, they are not easy to extend to other tasks. In practice, Cao *et al.*[4] build a model that contains two branches: one for the prediction of keypoints' heatmaps and the other for pairwise relationship. Although this model runs in real time, its result is not accurate and only for a single task. Papandreou *et al.*[5] use a convolutional network to detect keypoints and predict their relative displacement. For the segmentation task, they propose a part-induced geometric embedding descriptor which associates semantic person pixels with their corresponding person instance. However, a key limitation of their method for segmentation is its reliance on keypoint-level annotations for training. With a lot of refinement, their method still fails to achieve a satisfactory result.

In contrast, the top-down approaches [6, 7, 8] are pop-

978-1-5386-9552-4/19/$31.00 ©2019 IEEE
DOI 10.1109/ICME.2019.00188

ular as they offer precision and have the potential to extend to other tasks. Papandreou *et al.*[7] use ResNet [9] with dilated convolutions [10] which has been successful in the segmentation task. They adopt a combined classification and regression approach to model a keypoint's location as a disk-shaped mask and use 2D-offset vector fields to accurately locate keypoints. However, this method must strictly follow the 2-stage pipeline: using an off-the-shelf person detector first [11, 12] and then running a single-person pose estimation [13]. Furthermore, the scope of a disk is an empirical value which is unsuitable in unconstrained environment and the whole framework is not flexible for other tasks. He *et al.* [8] combine instance segmentation and pose estimation in a unified framework which supports end-to-end training. Instead of a disk-shaped mask, they use one-hot mask for keypoints prediction. Although the model is flexible, simple and fast, it still fails to meet the requirement of accuracy in a particular scenario compared with other top-down methods.

In this paper, we revisit the excellent top-down method for human pose estimation and adjust it to be surprisingly effective and flexible. Our system, called anchor-center based person network, achieves the comparable state-of-the-art results on COCO keypoints task and the state-of-the-art accuracy on COCO person instance segmentation task. Inspired by the Mask R-CNN framework which can do multi-tasks parallelly, we extend the Faster R-CNN by adding two well-designed branches: one for instance-specific poses and the other for mask prediction. Instead of using a person-detector, we use the person proposal network to predict the location and scale of anchors. The person proposal network enables our system to train and test in an end-to-end manner. Combining the annotation of bounding boxes, we model keypoints locations as anchor-center based labels and adopt 2D-offset vector fields to get a more precise estimation. The design can help our system improve the accuracy of keypoints localization and make association in complex and crowded multi-person scenes more accurate. Considering the uncontrollability of receptive fields in a deep CNN, we use ResNet-FPN [14] with deformable convolutions [15] and RoIAlign [8] to generate denser and aligned feature maps for modeling geometric transformations. The ablation experiments demonstrate that these well-designed modules can achieve high-quality results on the public benchmarks. On the whole, our system can simultaneously detect persons, make multi-person pose estimation, and segment person instances at the speed of 5 fps without any tricks.

2. OUR METHODS

Fig.3. illustrates the pipeline of our method. The input of our system is a color image with a fixed size $W \times H$. The output is the localization of a group of bounding boxes, scores, keypoints and masks for people in images. First, a feed-forward network with a powerful backbone is designed for feature extraction. Then the person proposal network generates a group of RoIs potentially contain person-instances. After a pixel-to-pixel aligned layer, three kinds of head branches for bounding-box regression, keypoints localization and mask prediction are applied separately on each RoI.

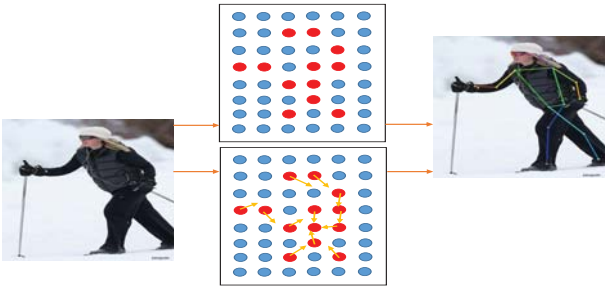

Fig. 2. Heatmaps and 2D-offset vector fields for pose estimation based on a disk-shaped mask label.

2.1. Person Proposal Network

The Faster-RCNN [11] has been widely used in object detection. It consists of two stages: the first stage is called Region Proposal Network (RPN), which is a fully-convolutional network that simultaneously predicts object proposals and object scores at each position. Like Mask R-CNN, we also inherit the advantage of Faster R-CNN to generate relatively high quality proposals. In this paper, region proposal and subsequent branches of our system have been trained using only the person category in the COCO dataset. Therefore, we name it Person Proposal Network (PPN). The output of PPN will provide a solid foundation for the subsequent tasks: person detection, human pose estimation and person instance segmentation. We use a tuple (A_w^i, A_h^i) to represent the height and width of the i-th anchor.

2.2. Person Pose Estimation

For each keypoint, we introduce a set of anchor-center based labels for training which also include the prediction of heatmap and 2D-offset vector fields. Why do not we use one-hot mask or a disk-shaped mask label for training? The following section will describe how to find an optimal design for our pipeline.

One-Hot Mask Label: It is obvious that the one-hot mask label is not very elegant in the complex vision tasks. In the Mask-RCNN, it models keypoints' localization as a one-hot mask to predict k masks, one for each of k keypoint types (*e.g.*,left shoulder, right hand). The training target is a one-hot $m \times m$ binary mask where only a single pixel is marked as a target. Compared with a disk-shaped mask label, it lacks the composition of the 2D-offset vectors which can get a more precise location of each keypoint.

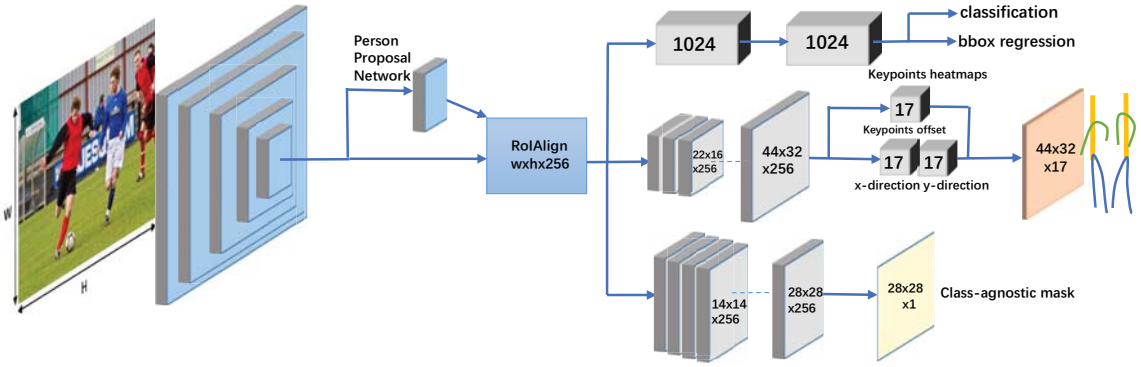

Fig. 3. The pipeline of our method consists of a powerful backbone and different head branches for multiple tasks.

A Disk-Shaped Mask Label: To address the above limitations, we formulate keypoints localization as a combined classification and regression problem. On each RoI, we traverse every pixel and determine whether it is within the scope of certain keypoint. The scope is often empirically given ($R = 25$ pixels in [7]). For a more precise result, we predict 2D-offset vectors in this scope. For each position x_i and keypoint k, we compute the probability $h_k(x_i) = 1$ if $\|x_i - l_k\| \leq R$, i.e., the point x_i is within a disk of radius R from the location l_k of keypoint k . We use s to represent the output stride and set it equals to 8 pixels. The size of the final outputs are 3D tensors with a fixed size $w \times h \times k$. We use a tuple (w^i, h^i) to represent the i-th heatmap's resolution and it is defined as

$$(w^i, h^i) = ((A^i_w, A^i_h) - 1)/s + 1. \tag{1}$$

Instead of using the bounding boxes cropped from raw images, we resize each proposal (RoI), candidates of person-instances, to a fixed size 353×257 before RoIAlign layer. Finally, the keypoint head branch with 17 1x1 convolution module outputs 3 channels: one channel for the heatmaps and the other two channels for the 2D-offset vectors.

Heatmaps: Each heatmap is a 3D tensor with a fixed size $w^i \times h^i \times k$. Different public benchmarks have different number of keypoints ($k = 15$ or $k = 17$). In some cases, heatmaps can also be called confidence maps. Because the value in each grid of heatmap represents the probability that a keypoint at the grid (h_x, h_y) has the corresponding type.

Offset Vectors: As explained above, heatmaps can only represent the approximate localization of the keypoints. To estimate the more precise positions of keypoints, we use the offset vectors as shown in Fig.2. Let $(l_{k,x}, l_{k,y})$ denote the offset vectors, and we can get the accurate position of keypoints by

$$(I_x, I_y) = (h_x, h_y)s + (l_{x.k}, l_{y,k}). \tag{2}$$

Some of our experiments show that one-hot mask label and a disk-shaped mask label can not perform well in complex multi-person scenes due to contact, occlusion and limb articulations. Based on the annotation of bounding boxes, we design a set of new labels for training to solve the limitation.

Anchor-Center based Label: For each keypoint, we formulate the prediction of its heatmap and offset vectors as a regression problem. Let B_i, $i = 1, ..., N$ denote the i-th bounding box annotation in an image. In COCO dataset [16], each person instance is labeled by a bounding box and there are $K = 17$ ground truth annotations for face and body parts. Let C_i be the center coordinates of B_i and G^*_i be the ground truth of keypoints in i-th bounnding box. The i-th bounding box corresponds to the i-th person instance. Then we can calculate the distances between the keypoints of each person instance and the center of corresponding bounding box so that a distance matrix D_i with size $N \times K$ is obtained, where entry $D_i[n, k]$ denotes the distance between the k-th keypoint of the i-th person instance and the center of n-th anchor.

For two adjacent anchor centers C_i and C_{i-1}, we calculate their distance in the x direction.The result is stored in a vector named $D_{x.\Delta c}$, which contains $n - 1$ elements. In the same way, we get a vector in the y direction named $D_{y,\Delta c}$. Then, we calculate the average of the two vectors mentioned above. Let $m_{x,\Delta c}$ denote the average of the $D_{x,\Delta c}$ and $m_{y.\Delta c}$ for $D_{y,\Delta c}$.

Like the definition in **A Disk-Shaped Mask Label**, we can get their mapping on the feature maps by:

$$(m_{x,f}, m_{y,f}) = (m_{x,\Delta c}, m_{y,\Delta c})/s. \tag{3}$$

After corresponding scaling in the x and y directions, we get a new value of the scope based on the anchor center,

$$R_f = R\sqrt{(m_{x,f})^2 + (m_{y,f})^2}. \tag{4}$$

Instead of predicting the 2D-offset vector fields mentioned above, we predict the distance of all keypoints based on the center of each bounding box shown in the Fig.1 as our new offset vectors. Finally, a 2D regression method is used

Fig. 4. The visualization results of multi-person pose estimation and person instance segmentation on COCO test set.

Table 4. The below four tables from left to right named (a), (b), (c), (d) represent our ablation study.

Keypoints Head	AP_{kps}
8-conv, 512-d	0.642
3-conv, 256-d	0.653

Solution	AP_{kps}
cls+reg	0.646
both reg	0.657

Task	AP_{kps}	AP_{mask}
kps-only	0.649	-
kps & mask	0.657	0.465

Conv	AP_{kps}	AP_{mask}
algin & none	0.652	0.455
align & deformable	0.657	0.465

to solve the new offset problem for each each position x_i and the k-th keypoint independently.

2.3. Person Instance Segmentation

Our method is based on parallel prediction of masks and class labels for person category, which is very flexible and accurate. And it can be easily extended to any category of instance segmentation. In all experiments, we use ResNet-FPN network as backbone. With deformable convolutions, we not only get more precise localization of keypoints but also more accurate segment proposals.

deformable convolution:The key to the segmentation problem is how to solve pixel-to-pixel alignment problem. Therefore,we use the RoIAlign layer introduced in Mask R-CNN which can help features to be well aligned. Considering the uncontrollability of the filters' view in deep convolutional neural network, it will have a certain impact on the final segmentation results. We employ deformable convolutions with unit equals to [0,1,1,3] and the number of groups in each unit equals to [0,4,4,3], which can control the field-of-view adaptively and model deformations of object more robustly (more details in [15]). As shown in experiments section, deformable convolution can lead to large improvements.

2.4. An End-to-End Solution

We report experimental results on a single machine with 4 TITAN XP GPUs. Mini-batch size is 4. The base learning rate is 1e-4 with warm-up step, momentum value set to 0.9 and weight decay equals to 4e-5. The learning rate drops to 1e-5 at 14 epochs and 1e-6 at 17 epochs. There are 18 epochs in total with stochastic gradient descent. Other hyper-parameters follow the Mask R-CNN apart from the special parameters mentioned above and well-designed structure in pose estimation and instance segmentation. Data augmentation includes scale $[-0.1, 0.2]$, rotation $[-40, 40]$ degrees and flip.

Training: Our loss function on each RoI is defined as:

$$L = L_{cls} + L_{box} + L_{hm} + L_{offset} + L_{mask}, \tag{5}$$

L_{cls} and L_{box} have an identical definition as Faster-RCNN (more details can be found in [11]).

As for the localization of keypoints, we use a well-designed keypoints branch on the top of our backbone with two convolutional output heads. For each position x_i and each type of keypoint k, we use $\tilde{r}_{k,j}(x_i)$ to indicate whether the keypoint k exists in the region which mentioned in **Anchor-Center based Label**. The label for training is $\tilde{r}_{k,j}(x_i)$ which is a grid only contain zeros and ones, with $\tilde{r}_{k,j}(x_i) = 1$ if $\|x_i - C_j\| \leq R_f$ and 0 otherwise. Therefore, the L_{hm} can be represented as the sum of robust loss function (smooth L_1 [11]) for each position and keypoint separately. For the solution of offset head, we also treat it as a regression problem. The corresponding loss L_{offset} can also be represented as the sum of smooth L_1 for the predicted and ground truth offset vectors. We only penalize the loss that the position x_i is in the scope of R_f.

With these definitions, our loss function for human pose estimation in images is defined as:

$$L_{hm} = \frac{1}{k \times \omega} \sum_i L_s(H_i, H_i^*) \tag{6}$$

,

$$L_{offset} = \frac{1}{k \times \omega} \sum_i L_s(O_i, O_i^*) \tag{7}$$

where ω is the batch size, H_i is the prediction of heatmap, H_i^* is the ground truth of heatmap, O_i is the prediction of offset

vectors, O_i^* is the ground truth of offset vectors and L_s is the smooth L1 loss function. Since our system only cares whether a person instance exists on each RoI, we do not implement the branch which generates masks for every class (80 categories in COCO). Therefore, we take advantage of the class-agnostic masks [8] rather than class-specific masks. The reason why we use class-agnostic masks is that our system is based on parallel prediction of masks and instance category, which decouples segmentation and classification to a large extend. We use the cross entropy loss on each sampled RoI. So the loss function for person instance segmentation is defined as:

$$L_{mask} = -\sum_i G_m^*(x_i) log(G_m(x_i)) \tag{8}$$

Here, for each position x_i, $G_m^*(x_i)$ is the labeled mask information and $G_m(x_i)$ is the output of our mask branch.

3. EXPERIMENTAL EVALUATION

3.1. Experimental Setup

We train the model and evaluate our method on COCO benchmark about keypoints challenge and instance segmentation only for the person category. Our backbone models were pretrained on ImageNet dataset [17]. Our training dataset contains 64115 images and validation dataset contains 2693 images with person instance. We also evaluate the system performance on the test-dev split which contains 20288 images.

For a fair comparison with the state-of-the-art results, we only use COCO 2017 training dataset for training and COCO 2017 validation dataset and test-dev split for evaluating.

3.2. COCO multi-person pose estimation results

Table 1 shows the performance comparison on the COCO keypoint test-dev split. We evaluate the person keypoint AP (AP^{kp}) using ResNet-50-FPN backbone. Because our method makes up for the defects of one-hot label used in Mask R-CNN. Our single-scale inference results have exceeded the original results of Mask R-CNN. Our results also outperform some very popular multi-peron pose estimation: Openpose and Associating Embedding. In terms of speed, we have no way to exceed the bottom-up methods. However, we can do multi-task in parallel with the well-designed branches. Compared with a single task, multi-task collaborative training can also improve the accuracy of each task at the same time. The experimental results shows our method achieves comparable results on the COCO keypoints task.

3.3. COCO person instance segmentation results

Without any other tricks, we only use the ResNet-50-FPN with deformable convolutions to generate denser feature maps and multi-task collaborative training. The output stirde = 8. Table 2 and 3 show our person instance segmentation results

Table 1. Performance on COCO keypoints test-dev split.

	AP	$AP^{0.50}$	$AP^{0.75}$	AP^M	AP^L
Bottom-up methods:					
Openpose [4] (+refine)	0.618	0.849	0.675	0.571	0.682
Assoc. Embed [3] (multi-scale)	0.630	0.857	0.689	0.580	0.704
PersonLab:					
ResNet-101 [5] (single-scale)	0.655	0.871	0.714	0.613	0.715
Top-down methods:					
Mask R-CNN [8]	0.631	0.873	0.687	0.578	0.714
G-RMI [7] COCO-only(ResNet-101)	0.649	0.855	0.713	0.623	0.700
ours:					
ResNet-50-FPN(single-scale)	**0.644**	0.890	0.708	0.602	0.730
ResNet-50-FPN(multi-scale)	**0.647**	0.895	0.711	0.605	0.739

Table 2. Performance on COCO person instance segmentation test-dev split.

	AP	$AP^{0.50}$	$AP^{0.75}$	AP^S	AP^M	AP^L
FCIS [18] (baseline)	0.334	0.641	0.318	0.090	0.411	0.618
FCIS [18] (multi-scale)	0.386	0.693	0.410	0.164	0.481	0.621
PersonLab [5]:						
ResNet-101 (single-scale)	0.377	0.659	0.394	0.166	0.480	0.595
ResNet-101 (multi-scales)	0.411	0.686	0.445	0.215	0.496	0.626
ours(single-scale):						
ResNet-50-FPN	**0.457**	0.745	0.435	0.239	0.510	0.648

on COCO test-dev and val-dev. As shown in the Table 3, our system has outperformed the Mask R-CNN, which indicates that the proposed technique is very effective in person instance segmentation. In Table 2, compared with all the person instance segmentation methods to be either box-free or box-based, our person instance segmentation results surpass theirs without bells and whistles. Especially, our results far outperforms the more complex FCIS [18].

3.4. Ablation Study:

Ablation studies are used to validate the performance of our well-designed modules. We report abliations on the remaining 5k val split.

New Keypoints Head *vs.* Original Head: Table 4 (a) evaluates the effect of our system using new keypoints head. Compared with a stack of eight 3×3 512-d conv layers in Mask R-CNN. We only use a stack of three 3×3 256-d conv layers and get the comparable result(outperform Mask R-CNN) on COCO benchmark, which greatly reduce the amount of computation.

Regression *vs.* Classification + Regression problem: Table 4 (b) shows that we adopt a regression approach(anchor center based label) compared with a combined classification and regression approach (a-disk mask label:*cls for heatmap, reg for offsets*) to solve the location of keypoints. It has 0.646 kps AP *vs.* 0.657. This suggests that our improvement is effective.

Single task *vs.* Multi-task Learning: Table 4 (c) reports the single task (human pose estimation) benefits from the multi-task training (human pose estimation and instance segmentation, and 0.8% gain is obtained.

Deformable *vs.* No Deformable:Table 4 (d) evaluates the effect of using deformable convolution. It indicates that it is critical to consider a relatively large and adaptive filed-of-

Table 3. Performance on COCO person instance segmentation val split

	AP	$AP^{0.50}$	$AP^{0.75}$	AP^{S}	AP^{M}	AP^{L}
Mask R-CNN (ResNet-101-FPN)	0.455	0.798	0.472	0.239	0.511	0.611
PersonLab:						
ResNet-101 (single-scale)	0.382	0.661	0.397	0.164	0.476	0.592
ResNet-101 (multi-scales)	0.414	0.684	0.447	0.213	0.492	0.621
ours(single-scale):						
ResNet-50-FPN	**0.464**	0.800	0.485	0.255	0.517	0.674
ResNet-101-FPN	**0.465**	0.804	0.487	0.247	0.521	0.685

view in pixel-wise predictions task.

3.5. Qualitative Results

Figure 4 shows some qualitative results of our model for multi-person pose estimation and person instance segmentation on the COCO test set.

4. CONCLUSION

In this work, we efficiently and effectively address the problem of multi-person pose estimation and person instance segmentation in a flexible and end-to-end manner. We present a conceptually simple system, consisting of a person proposal network followed by three different head branches for person detection, human pose estimation, and instance segmentation. With a set of well-designed labels based on anchor-center and deformable convolutions, our method achieves the comparable results on challenging COCO person keypoint task and the state-of-the-art accuracy on COCO person instance segmentation task.

Acknowledgments

This work is supported by National Key RD Program of China under contract No. 2017YFB1002203, and NSFC Key Projects of International (Regional) Cooperation and Exchanges under Grant 61860206004.

5. REFERENCES

[1] L. Pishchulin, E. Insafutdinov, S. Tang, B. Andres, M. Andriluka, P. Gehler, and B. Schiele, "Deepcut: Joint subset partition and labeling for multi person pose estimation," *In CVPR*, 2016.

[2] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multiperson pose estimation model," *In ECCV*, 2016.

[3] Newell. A, Huang. Z, and Deng. J, "Associative embedding: End-to-end learning for joint detection and grouping," *In NIPS*, 2017.

[4] Cao. Z, Simon. T, Wei. S.E, and Sheikh. Y, "Realtime multi-person 2d pose estimation using part affinity fields," *In CVPR*, 2017.

[5] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy, "Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model," *In ECCV*, 2018.

[6] Newell. A, Yang. K, and Deng. J, "Stacked hourglass networks for human pose estimation," *In ECCV*, 2016.

[7] Papandreou. G, Zhu. T, Kanazawa. N, Toshev. A, Tompson. J, Bregler. C, and Murphy. K, "Towards accurate multi-person pose estimation in the wild," *In CVPR*, 2017.

[8] K. He, Gkioxari. G, P. Dollr, and R. Girshick, "Mask r-cnn," *In ICCV*, 2017.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *In CVPR*, 2016.

[10] Chen. L.C, Papandreou. G, Kokkinos. I, Murphy. K, and Yuille. A.L, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs.," *In IEEE Transaction on Pattern Analysis and Machine Intelligence*, 2017.

[11] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *In NIPS*, 2015.

[12] Jifeng Dai, Yi Li, K. H, and J Sun, "R-fcn: Object detection via region-based fully convolutional networks," *In NIPS*, 2016.

[13] Dalal. N and Triggs. B, "Histograms of oriented gradients for human detection," *In CVPR*, 2005.

[14] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," *In CVPR*, 2017.

[15] Dai. J, Qi. H, Xiong. Y, Li. Y, Zhang. G, Hu. H, and Wei. Y, "Deformable convolutional networks," *In ICCV*, 2017.

[16] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft coco: Common objects in context," *In ECCV*, 2014.

[17] Krizhevsky. A, Sutskever. I, and Hinton. G. E, "Imagenet classification with deep convolutional neural networks," *In NIPS*, 2012.

[18] Li. Y, Qi. H, Dai. J, Ji. X, and Wei. Y, "Fully convolutional instance-aware semantic segmentation," *In CVPR*, 2017.