Multi-Person 2D Pose Estimation

First Author
Institution1
Institution1 address
firstauthor@i1.org

Second Author
Institution2
First line of institution2 address
secondauthor@i2.org

Abstract

The topic of multi-person pose estimation has been largely improved recently, especially with the development of convolutional neural network. In this report, we want to compare the most popular methods in these years, select some methods to implement to find the pros and cons of these methods, and see if there are any space to get promotion or application scenarios. We also try to transfor NMS to bottom-up approach to detect and remove the redundant poses, which can promote the precision of CMU Pose.

1. Introduction

Multi-person pose estimation is to recognize and locate the keypoints for all persons in the image, which is a fundamental research topic for many visual applications like human action recognition and human-computer interaction.

Inferring the pose of multiple people in images, especially socially engaged individuals, presents a unique set of challenges. First, each image may contain an unknown number of people that can occur at any position or scale. Second, interactions between people induce complex spatial interference, due to contact, occlusion, and limb articulations, making association of parts difficult. Third, runtime complexity tends to grow with the number of people in the image, making realtime performance a challenge.

Recently, the problem of multi-person pose estimation has been greatly improved by the involvement of deep convolutional neural networks [15, 22]. For example, in [4], convolutional pose machine is utilized to locate the keypoint joints in the image and part affinity fields (PAFs) is proposed to assemble the joints to different person. Mask-RCNN [14] predicts human bounding boxes first and then warps the feature maps based on the human bounding boxes to obtain human keypoints. Cascaded Pyramid Network (CPN) [7] propose a novel network structure to address these hard joints such as occluded keypoints, invisible keypoints and crowded background, which cannot be well localized.

Based on this task, there are two different pipelines called top-down and bottom-up. For top-down method, human detector is first adopted to generate a set of human bounding boxes, followed by specific network for keypoint localization in each human bounding box. For bottom-up method, we need to detect all the keypoints first, and then consider to associate them as natural poses like the human do. In this report, we will introduce the development of the two methods and implement the representation of these two approaches to see the advantages and disadvantages of them, at the last, we try to transfor NMS to bottom-up approach to improve its precision.

2. Related Work

Human pose estimation is an active research topic for decades. Classical approaches tackling the problem of human pose estimation mainly adopt the techniques of pictorial structures [10, 1] or graphical models [6]. More specifically, the classical works [1, 32, 11, 31, 8, 43, 28, 18] formulate the problem of human keypoints estimation as a tree-structured or graphical model problem and predict keypoint locations based on hand-crafted features. Recent works [26, 12, 3, 17, 41, 42] mostly rely on the development of convolutional neural network(CNN) [22, 15], which largely improve the performance of pose estimation. In this paper, we mainly focus on the methods based on the convolutional neural network. The topic is categorized as single-person pose estimation that predicts the human keypoints based on the cropped image given bounding box, and multi-person pose estimation that require further recognition of the full body poses of all persons in one image.

2.1. Single Person Pose Estimation

In single person pose estimation, the pose estimation problem is simplified by only attempting to estimate the pose of a single person, and the person is assumed to dominate the image content. Conventional methods considered pictorial structure models. For example, tree models [40, 33, 44, 39] and random forest models [35, 8] have demonstrated to be very efficient in human pose estimation.

Graph based models such as random field models [19] and dependency graph models have also been widely investigated in the literature [13, 36, 21, 29].

More recently, deep learning has become a promising technique in object/face recognition, and human pose estimation is of no exception. Toshev *et al*. firstly introduce CNN to solve pose estimation problem in the work of DeepPose [38], which proposes a cascade of CNN pose regressors to deal with pose estimation. Tompson *et al*. [37] attempt to solve the problem by predicting heatmaps of keypoints using CNN and graphical models. Later works such as Wei *et al*. [41] and Newell *et al*. [26] show great performance via generating the score map of keypoints using very deep convolutional neural networks. Wei *et al*. [41] propose a multi-stage architecture, i.e., first generate coarse results, and continuously refine the result in the following stages. Newell *et al*. [26] propose an U-shape network, i.e., hourglass module, and stack up several hourglass modules to generate prediction. Carreira *et al*. [5] uses iterative error feedback to get pose estimation and refine the prediction gradually. Lifshitz *et al*. [23] uses deep consensus voting to vote the most probable location of keypoints. Gkioxary *et al*. [12] and Zisserman *et al*. [2] apply RNN-like architectures to sequentially refine the results.

2.2. Multi-Person Pose Estimation

Multi-person pose estimation is gaining increasing popularity recently because of the high demand for the real-life applications. However, multi-person pose estimation is challenging owing to occlusion, various gestures of individual persons and unpredictable interactions between different persons. The approach of multi-person pose estimation is mainly divided into two categories: bottom-up approaches and top-down approaches.

2.2.1 Bottom-Up Approaches

Bottom-up approaches [4, 25, 30, 17] directly predict all keypoints at first and assemble them into full poses of all persons. DeepCut [30] interprets the problem of distinguishing different persons in an image as an Integer Linear Program (ILP) problem and partition part detection candidates into person clusters. Then the final pose estimation results are obtained when person clusters are combined with labeled body parts. DeeperCut [17] improves DeepCut [30] using deeper ResNet [15] and employs image-conditioned pairwise terms to get better performance. Zhe Cao *et al*. [4] map the relationship between keypoints into part affinity fields (PAFs) and assemble detected keypoints into different poses of people. Newell *et al*. [25] simultaneously produce score maps and pixel-wise embedding to group the candidate keypoints to different people to get final multi-person pose estimation.

2.2.2 Top-Down Approaches

Top-down approaches [27, 16, 14, 9, 7] interpret the process of detecting keypoints as a two-stage pipeline, that is, firstly locate and crop all persons from image, and then solve the single person pose estimation problem in the cropped person patches. Papandreou *et al*. [27] predict both heatmaps and offsets of the points on the heatmaps to the ground truth location, and then uses the heatmaps with offsets to obtain the final predicted location of keypoints. Mask-RCNN [14] predicts human bounding boxes first and then crops the feature map of the corresponding human bounding box to predict human keypoints. If top-down approach is utilized for multi-person pose estimation, a human detector as well as single person pose estimator is important in order to obtain a good performance.

3. Methods

In this report, we choose two classical methods: CMU-Pose [4], the winner of the COCO 2016 keypoint challenge to represent the bottom-up approach and Cascaded Pyramid Network [7], the winner of the COCO 2017 keypoint challenge to represent the top-down approach.

3.1. CMU-Pose

This method presents the first bottom-up representation of association scores via Part Affinity Fields (PAFs), a set of 2D vector fields that encode the location and orientation of limbs over the image domain. It demonstrates that simultaneously inferring these bottom-up representations of detection and association encode global context sufficiently well to allow a greedy parse to achieve high-quality results, at a fraction of the computational cost. And it has publically presented the first realtime system for multi-person 2D pose detection.

Figures 1 illustrates the overall pipeline of this method. The system takes, as input, a color image of size $w \times h$ (Figures 1a) and produces, as output, the 2D locations of anatomical keypoints for each person in the image (Figures 1e). First, a feedforward network simultaneously predicts a set of 2D confidence maps $\mathbf{S}$ of body part locations (Figures 1b) and a set of 2D vector fields $\mathbf{L}$ of part affinities, which encode the degree of association between parts (Figures 1c). The set $\mathbf{S} = (\mathbf{S}_1, \mathbf{S}_2, ..., \mathbf{S}_J)$ has J confidence maps, one per part, where $S_j \in \mathbb{R}^{w \times h}$, $j \in \{1...J\}$. The set $\mathbf{L} = (\mathbf{L}_1, \mathbf{L}_2, ..., \mathbf{L}_C)$ has C vector fields, one per limb, where $\mathbf{L}_c \in \mathbb{R}^{w \times h \times 2}$, $c \in \{1...C\}$, each image location in $\mathbf{L}_c$ encodes a 2D vector. Finally, the confidence maps and the affinity fields are parsed by greedy inference (Figures 1d) to output the 2D keypoints for all people in the image.

The architecture, shown in Figures 2, simultaneously predicts detection confidence maps and affinity fields that

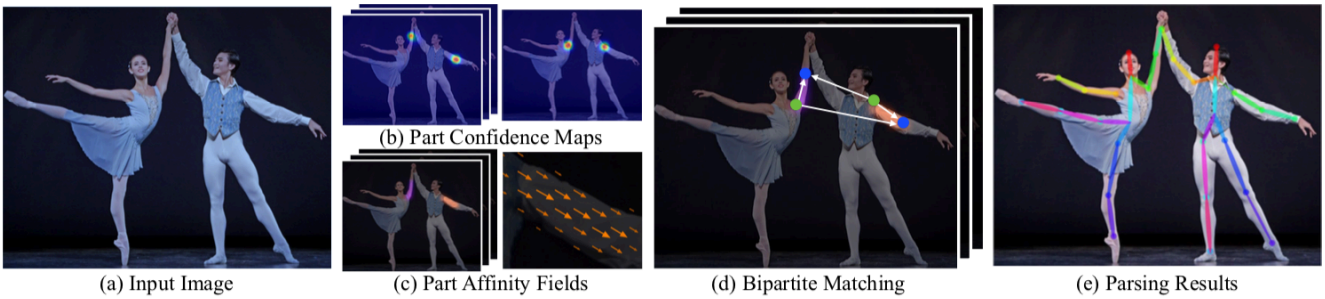

Figure 1. Overall pipeline.

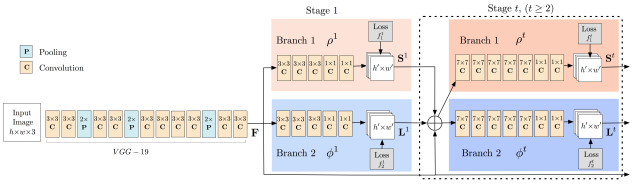

Figure 2. Architecture of the two-branch multi-stage CNN.

en- code part-to-part association. The network is split into two branches: the top branch, shown in beige, predicts the confidence maps, and the bottom branch, shown in blue, predicts the affinity fields. Each branch is an iterative prediction architecture, follow Wei *et al*. [41], which refines the predictions over successive stages, $t \in \{1, ..., T\}$, with intermediate supervision at each stage.

The image is first analyzed by a convolutional network (initialized by the first 10 layers of VGG-19 [34] and fine-tuned), generating a set of feature maps $\mathbf{F}$ that is input to the first stage of each branch. At the first stage, the network produces a set of detection confidence maps $\mathbf{S}^1 = \rho^1(\mathbf{F})$ and a set of part affinity fields $\mathbf{L}^1 = \phi^1(\mathbf{F})$, where ρ^1 and ϕ^1 are the CNNs for inference at Stage 1. In each subsequent stage, the predictions from both branches in the previous stage, along with the original image features $\mathbf{F}$, are concatenated and used to produce refined predictions,

$$\mathbf{S}^t = \rho^t(\mathbf{F}, \mathbf{S}^{t-1}, \mathbf{L}^{t-1}), \forall t \geq 2, \quad (1)$$

$$\mathbf{L}^t = \phi^t(\mathbf{F}, \mathbf{S}^{t-1}, \mathbf{L}^{t-1}), \forall t \geq 2, \quad (2)$$

where ρ^t and ϕ^t are the CNNs for inference at Stage t.

3.2. Cascaded Pyramid Network

This method presents a novel network structure called Cascaded Pyramid Network (CPN) which targets to relieve the problem from "hard" keypoints. More specifically, this algorithm includes two stages: GlobalNet and RefineNet. GlobalNet is a feature pyramid network which can successfully localize the "simple" keypoints like eyes and hands but may fail to precisely recognize the occluded or invisible keypoints. RefineNet tries explicitly handling the hard keypoints by integrating all levels of feature representations from the GlobalNet together with an online hard keypoint mining loss. In general, to address the multi-person pose estimation problem, a top-down pipeline is adopted to first generate a set of human bounding boxes based on a detector, followed by CPN for keypoint localization in each human bounding box.

3.2.1 GlobalNet

Here, we describe the network structure based on the ResNet backbone. We denote the last residual blocks of different conv features conv25 as C_2, C_3, ..., C_5 respectively. 3×3 convolution filters are applied on C_2, ..., C_5 to generate the heatmaps for keypoints. The shallow features like C_2 and C_3 have the high spatial resolution for localization but low semantic information for recognition. On the other hand, deep feature layers like C_4 and C_5 have more semantic information but low spatial resolution due to strided convolution (and pooling). Thus, usually an U-shape structure is integrated to maintain both the spatial resolution and semantic information for the feature layers. More recently, FPN [24] further improves the U-shape structure with deeply supervised information. We apply the similar feature pyramid structure for our key- points estimation. Slightly different from FPN, we apply 1×1 convolutional kernel before each element-wise sum procedure in the upsampling process. We call this structure as GlobalNet and an illustrative example can be found in Figure 3.

3.2.2 RefineNet

Based on the feature pyramid representation generated by GlobalNet, we attach a RefineNet to explicitly address the "hard" keypoints. In order to improve the efficiency and keep integrity of information transmission, our RefineNet transmits the information across different levels and finally integrates the informations of different levels via upsampling and concatenating as HyperNet [20]. Different from the refinement strategy like stacked hourglass [26], our RefineNet concatenates all the pyramid features rather than

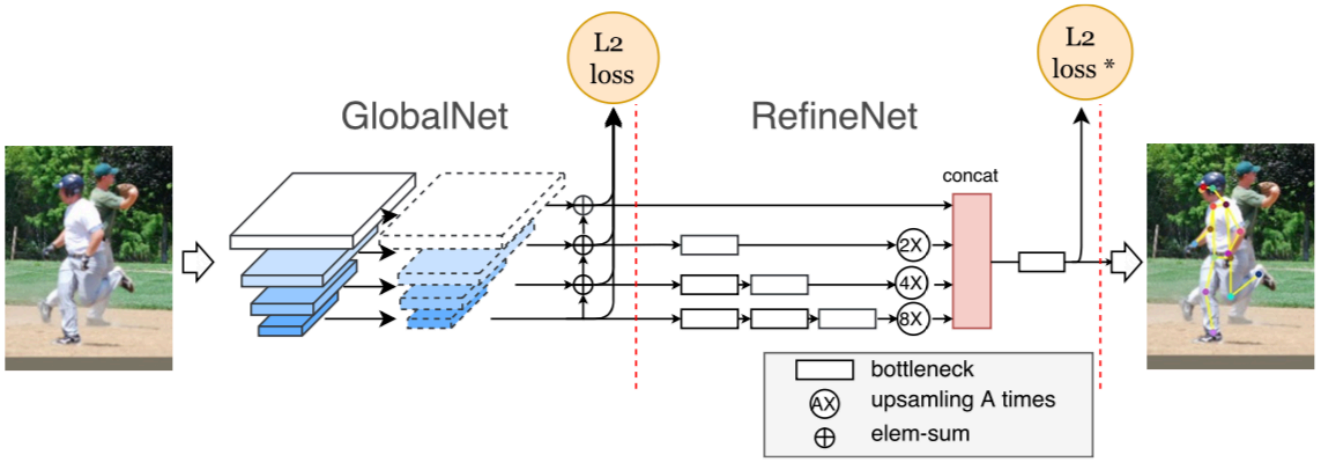

Figure 3. Cascaded Pyramid Network. "L2 loss*" means L2 loss with online hard keypoints mining.

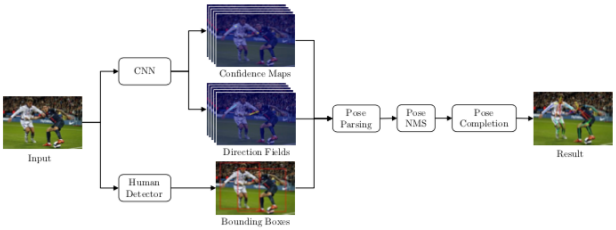

Figure 4. Overview. Given an image, firstly, estimate the confidence maps and the direction fileds by the trained CNN. Secondly, parse the pose for each person within corresponding bounding box. Thirdly, remove the redundant poses by pose NMS. Finally, complete the pose by associating the disconnected points.

simply using the upsampled features at the end of hourglass module. In addition, we stack more bottleneck blocks into deeper layers, whose smaller spatial size achieves a good trade-off between effectiveness and efficiency.

Here we discuss the losses used in our network. In detail, the loss function of GlobalNet is L2 loss of all annotated keypoints while the second stage tries learning the hard keypoints, that is, we only punish the top M ($M < N$) keypoint losses out of N(the number of annotated keypoints in one person, say 17 in COCO dataset). For M = 8, the performance of second stage achieves the best result for the balanced training between hard keypoints and simple keypoints.

3.3. Improvement

Inspired by top-down approach, we get an idea that we can use an operation like NMS in detection task to make the CMU-Pose more precious. The Figure 4 shows the overview.

3.3.1 Pose NMS

We try to transfor NMS from top-down approach into bottom-up approach to improve its precision, which is called Pose NMS to detect and remove the redundant poses. Firstly, select the most confident pose as the reference pose $\mathbf{Y}'$, then the other poses close to $\mathbf{Y}'$ are subject to elimination by applying an elimination criterion. Repeat this process on all of the poses set, until at most one unique pose remains in one bounding box.

Pose confidence. We define the pose confidence by taking into account the covering area of the pose, the confidence of the joints, and the confidence of the connections. Consider a pose $\mathbf{Y}$ with $\mathbf{J}$ joints: $\mathbf{Y} = (\mathbf{Y_1},...,\mathbf{Y_J})$, where $\mathbf{Y_j}$ is the location of joint j. Let $s_1(\mathbf{Y})$ and $s_2(\mathbf{Y})$ are the arithmetical average confidence score of the joints and the connections for pose Y. Then we define the confidence of pose $\mathbf{Y}$ as

$$Conf(\mathbf{Y}) = \alpha s_1(\mathbf{Y}) + \beta s_2(\mathbf{Y}) + \gamma \frac{S(B'(\mathbf{Y}))}{S(B(\mathbf{Y}))}, \quad (3)$$

where $B'(\mathbf{Y})$ is the minimum bounding box of $\mathbf{Y}$, $B(\mathbf{Y})$ is the bounding box of $\mathbf{Y}$, and $S(\cdot)$ denotes the area of the bounding box.

Elimination criterion. In order to eliminate the poses close to the reference pose $\mathbf{Y}$, we define a distance metric d($\mathbf{Y}$, $\mathbf{Y}'$) to measure the pose similarity,

$$d(\mathbf{Y}, \mathbf{Y}') = \frac{\sum_{j=1}^{J}(\mathbf{Y_j} \neq \mathbf{Y'_j})}{max(n_{\mathbf{Y}}, n_{\mathbf{Y}'})}, \quad (4)$$

where $n_{\mathbf{Y}}$ and $n_{\mathbf{Y}'}$ count the number of visible joints in $\mathbf{Y}$ and $\mathbf{Y}'$, $d(\mathbf{Y}, \mathbf{Y}')$ computes the percentage of unmatched joints between $\mathbf{Y}$ and $\mathbf{Y}'$. We specify a threshold η as the elimination criterion,

AP/AR	IoU	area	maxDets	value
AP	0.50:0.95	all	20	61.8
AP	0.50	all	20	84.9
AP	0.75	all	20	67.5
AP	0.50:0.95	medium	20	57.1
AP	0.50:0.95	large	20	68.2
AR	0.50:0.95	all	20	66.5
AR	0.50	all	20	87.2
AR	0.75	all	20	71.8
AR	0.50:0.95	medium	20	60.6
AR	0.50:0.95	large	20	74.6

Table 1. CMU-Pose

AP/AR	IoU	area	maxDets	value
AP	0.50:0.95	all	20	72.1
AP	0.50	all	20	91.4
AP	0.75	all	20	80.0
AP	0.50:0.95	medium	20	68.7
AP	0.50:0.95	large	20	77.2
AR	0.50:0.95	all	20	78.5
AR	0.50	all	20	95.1
AR	0.75	all	20	85.3
AR	0.50:0.95	medium	20	74.2
AR	0.50:0.95	large	20	84.3

Table 2. Cascaded Pyramid Network

$$f(\mathbf{Y}, \mathbf{Y}|\eta) = \mathbf{1}(d(\mathbf{Y}, \mathbf{Y}') \leq \eta) \tag{5}$$

if $d(\cdot)$ is smaller than , the output of $f(\cdot)$ is 1, which indicates that pose $\mathbf{Y}$ should be eliminated due to redundancy with respect to the reference pose $\mathbf{Y}'$.

3.3.2 Pose Completion

Pose completion aims to associate disconnected joints caused by truncation or heavy occlusion to the corresponding pose. Motivated by single-person pose estimation methods, we take a very simple rule: for each missing joint in pose $\mathbf{Y}$, we find a point with the highest confidence in the corresponding cropped confidence map. If this point has already been associated to another pose, we find the next highest point constantly until it has not been associated to any other poses. Then we add this point to pose $\mathbf{Y}$.

4. Experiment

In section, we simply implement these two methods and run the codes on COCO 2017 keypoint challenge. The following tables are the results.

The comparision shows that the mAP of Cascaded Pyramid Network is much higher than CMU-Pose, one reason is that the Cascaded Pyramid Network uses top-down approach, which needs to first use the human detector to find all the bounding boxes of the peopele in one image, and then crop and use single person estimation, this truely gets better precision. But on the other hand, it will cost more time to deal with one image and hardly achieve realtime detection, this is why CMU-Pose is more practical in real life. The method has achieved the speed of 8.8 fps for a video with 19 people on a laptop with one NVIDIA GeForce GTX-1080 GPU.

	AP	AP50	AP75	APM	APL
CPN	72.1	91.4	80.0	68.7	77.2
CMU-Pose	61.8	84.9	67.5	57.1	68.2
Improvement	66.0	94.8	72.9	52.9	67.8

Table 3. Improvement method.

Table 3 shows the NMS truly improves the average precision of CMU Pose, but the speed slow down because of the NMS operation.

5. Conclusion

In this report, we introduce the challange of human pose estimation and talk about the bottom-up and top-down approachs. We use the CMU-Pose and the Cascaded Pyramid Network as representation to talk about the advantages and disadvantages of the two approachs. From the experiment results, we can see the precision and the speed is still not good enough to use in industrial products. How to get better and quicker results is still a big challange for us.

References

[1] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1014–1021. IEEE, 2009.

[2] V. Belagiannis and A. Zisserman. Recurrent human pose estimation. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 468–475. IEEE, 2017.

[3] A. Bulat and G. Tzimiropoulos. Human pose estimation via convolutional part heatmap regression. In *European Conference on Computer Vision*, pages 717–732. Springer, 2016.

[4] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1611.08050*, 2016.

[5] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik. Human pose estimation with iterative error feedback. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4733–4742, 2016.

[6] X. Chen and A. L. Yuille. Articulated pose estimation by a graphical model with image dependent pairwise relations. In *Advances in neural information processing systems*, pages 1736–1744, 2014.

[7] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun. Cascaded pyramid network for multi-person pose estimation. *arXiv preprint arXiv:1711.07319*, 2017.

[8] M. Dantone, J. Gall, C. Leistner, and L. Van Gool. Human pose estimation using body parts dependent joint regressors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3041–3048, 2013.

[9] H. Fang, S. Xie, and C. Lu. Rmpe: Regional multi-person pose estimation. pages 2353–2362, 2016.

[10] M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on computers*, 100(1):67–92, 1973.

[11] G. Gkioxari, P. Arbelaez, L. Bourdev, and J. Malik. Articulated pose estimation using discriminative armlet classifiers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3342–3349, 2013.

[12] G. Gkioxari, A. Toshev, and N. Jaitly. Chained predictions using convolutional neural networks. In *European Conference on Computer Vision*, pages 728–743. Springer, 2016.

[13] A. Gupta, T. Chen, F. Chen, D. Kimber, and L. S. Davis. Context and observation driven latent variable model for human pose estimation. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[14] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] S. Huang, M. Gong, and D. Tao. A coarse-fine network for keypoint localization. In *The IEEE International Conference on Computer Vision (ICCV)*, volume 2, 2017.

[17] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *European Conference on Computer Vision*, pages 34–50. Springer, 2016.

[18] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In *Computer vision and pattern recognition (CVPR), 2011 IEEE conference on*, pages 1465–1472. IEEE, 2011.

[19] M. Kiefel and P. V. Gehler. Human pose estimation with fields of parts. In *European Conference on Computer Vision*, pages 331–346. Springer, 2014.

[20] T. Kong, A. Yao, Y. Chen, and F. Sun. Hypernet: Towards accurate region proposal generation and joint object detection. In *Computer Vision and Pattern Recognition*, pages 845–853, 2016.

[21] L. Ladicky, P. H. Torr, and A. Zisserman. Human pose estimation using a joint pixel-wise and part-wise formulation. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3578–3585, 2013.

[22] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[23] I. Lifshitz, E. Fetaya, and S. Ullman. Human pose estimation using deep consensus voting. In *European Conference on Computer Vision*, pages 246–260. Springer, 2016.

[24] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. In *CVPR*, volume 1, page 4, 2017.

[25] 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*, pages 2277–2287, 2017.

[26] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016.

[27] 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 *CVPR*, volume 3, page 6, 2017.

[28] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Poselet conditioned pictorial structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 588–595, 2013.

[29] L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele. Strong appearance and expressive spatial models for human pose estimation. In *Proceedings of the IEEE international conference on Computer Vision*, pages 3487–3494, 2013.

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

[31] B. Sapp, C. Jordan, and B. Taskar. Adaptive pose priors for pictorial structures. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 422–429. IEEE, 2010.

[32] B. Sapp and B. Taskar. Modec: Multimodal decomposable models for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3681, 2013.

[33] B. Sapp, A. Toshev, and B. Taskar. Cascaded models for articulated pose estimation. In *European conference on computer vision*, pages 406–420. Springer, 2010.

[34] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[35] M. Sun, P. Kohli, and J. Shotton. Conditional regression forests for human pose estimation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3394–3401. IEEE, 2012.

[36] M. Sun, M. Telaprolu, H. Lee, and S. Savarese. An efficient branch-and-bound algorithm for optimal human pose estimation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1616–1623. IEEE, 2012.

[37] J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In *Advances in neural information processing systems*, pages 1799–1807, 2014.

[38] 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*, pages 1653–1660, 2014.

[39] F. Wang and Y. Li. Beyond physical connections: Tree models in human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 596–603, 2013.

[40] Y. Wang and G. Mori. Multiple tree models for occlusion and spatial constraints in human pose estimation. In *European Conference on Computer Vision*, pages 710–724. Springer, 2008.

[41] 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*, pages 4724–4732, 2016.

[42] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang. Learning feature pyramids for human pose estimation. In *The IEEE International Conference on Computer Vision (ICCV)*, volume 2, 2017.

[43] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1385–1392. IEEE, 2011.

[44] X. Zhang, C. Li, X. Tong, W. Hu, S. Maybank, and Y. Zhang. Efficient human pose estimation via parsing a tree structure based human model. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1349–1356. IEEE, 2009.