Paper:

Cycle Skeleton Structure for Occluded Multi-person 2D Pose Estimation

Longsheng Wei*[1,2,3], Xuefu Yu[1,2,3],Yuyang Ye[1,2,3], and Dapeng Luo[4]

[1] Key Laboratory of Geological Survey and Evaluation of Ministry of Education, Wuhan, China
[2] School of Automation, China University of Geosciences, Wuhan, China
[3] Hubei key Laboratory of Advanced Control and Intelligent Automation for Complex Systems, Wuhan, China
[4] School of Mechanical Engineering and Electronic Information
* E-mail: weilongsheng@cug.edu.cn

Abstract.There are two main pipelines in multi-person pose estimation. Compared to top-down approaches, bottom-up approaches save more computational cost in inference phase, but get lower accuracy for the final prediction result. Openpose is the first bottom-up approach and makes great progress in bottom-up field. However, this approach has room for improvement in both speed and accuracy. In this paper we modify the encoding method that uses only one heatmap to represent one connection to increase speed of inference step. And for tackling the isolated human parts problem caused by occlusion, we propose a new skeleton structure called Cycle Skeleton Structures for assembling step. In network structure, we use Hourglass module to extract multi-scale features at same time. In our experiment, we got accuracy improved on the subset of COCO validation dataset meanwhile speed up the runtime.

Keywords: Pose estimation, Bottom-up, Cycle skeleton structures

1. INTRODUCTION

With the great development of deep learning in computer vision field, we can move forward from classic image classification tasks towards more detailed visual understanding tasks such as pose estimation, instance segmentation in unconstrained environments. In this paper we tackle in the task of 2D multi-person pose estimation. Human pose estimation has some really practical applications and is heavily used in action recognition, animation, gaming, etc. For example, pose estimation is now utilized to analyse basketball player movements in some basketball matches. Given an image in unconstrained environment with multiple persons and, our goal is to identify every person instance, localize its facial and body keypoints, and assemble them to the right person pose. A host of computer vision applications, such as human-computer interaction, person and activity recognition, virtual or augmented reality, and sports supervising, can benefit from progress in these challenging tasks.

Recently there are two main approaches for tackling multi-person pose estimation. The top-down approaches [1,2] start by identifying and roughly localizing individual person instances through a bounding box of person object detector, and then perform single-person pose estimation algorithm in the region of the detected bounding box. In contrary, the bottom-up approaches [3,4], without person detector, first directly detect all the keypoints in a given wild image, and then assign them to corresponding person instance by keypoint assignment algorithm using the connection information among keypoints. Or assign keypoints via predicting each pixel's tag of which person or background belong to literature[5].

The Openpose approach[6] is the first bottom-up approach and makes great progress in bottom-up field. Our approach is mainly inspired by this approach and makes a little step forward. Therefore, the Openpose approach is selected to be the baseline approach in our paper.

During the process of reproducing the Openpose approach, we have found several problems that may be improved for practical application. The first problem is the method of encoding connection of two keypoints relating to each other. Two heatmaps are used to define one limb in the Openpose approach. To reduce the computation cost of inference step, we follow the encoding method of another state-of-the-art approach named Identity Mapping Hourglass Network (IMHN) [7], which uses only one heatmap to represent a connection. The second problem is that, when assembling the limb to full human pose, the Openpose approach uses a minimal number of edges to obtain a spanning tree skeleton of human pose instead of a complete connected graph. Although the Openpose approach simplifies the algorithm of assemble step, it will assign this part to wrong person if there is an isolated part. For tackling this problem, we propose a new skeleton structure for assembling step. And in network structure, we use Hourglass module [8] to extract multi-scale features ensuring that person and parts indifferent size can be detected at the same time.

We have made three contributions in our work: (1) we develop a new method called Improved Part Affinity Fields(IPAFs) to encode the connection of two adjacent keypoints, (2) we use a powerful network based on Hourglass module to generate multi scale heatmaps, and (3) we

adjust the skeleton graph to make pose assemble process robust.

2. RELATED WORK

2.1. Top-down approaches

Top-down approaches have got most of the state-of-the-art results on common keypoint detection dataset, such as HRNet [1], CPN [9], RMPE [10], G-RMI [11] and there are also works focusing on crowd scene pose estimation, like CrowdPose [12]. Benefiting from the existing well-trained person detectors, the state-of-the-art top-down approaches avoid the difficult subproblem of human body detection and can focus on single person pose estimation in the box region. However, the accuracy has been limited by the human detector heavily and two separate steps. The inference time will significantly increase if many persons appear in an image together.

2.2. Bottom-up approaches

The bottom-up approaches, e.g. MultiPoseNet [4], Associative embedding [5], Openpose [6], SMP [13], are more efficient in keypoint inference and do not rely on the human detector. However, they get lower accuracy in evaluation step. One main reason is that there are no perfect enough keypoint assignment algorithms for assembling keypoints to human pose. In pixel tag prediction-based method [5], it will cost the same computation resource as instance segmentation and has disadvantage in running speed. In connection-based method [6], in spite of speed advantage, there will be isolated parts occurs when some connections have not been detected. Recent connection-based methods have not dealed with this problem. Another main reason lies the fact that unlike the top-down method only take person box as input, input of bottom-up method is usually the full image, so that in heatmap the offset of only a few pixels away from the annotated keypoint location can lead to a big drop in the evaluation metrics on the MS-COCO benchmark. On the other side, the top-down approaches are immune to these challenges. And another reason which limits accuracy of bottom-up method accuracy is that persons and limbs have different sizes in the images which are difficult to detect at the only one time.

3. METHOD

We perform our approach in three steps: (1) we predict the keypoint heatmaps and connections of all persons in a given image, (2) we get the candidate keypoints and connections by performing Non-Maximum Suppression (NMS) on the inferred heatmaps, and (3) we perform the keypoint assignment algorithm and assemble all individual human poses using Cycle Skeleton Structure.

3.1. Definition of keypoint Heatmaps

We follow the Openpose approach to create joints heatmaps and connection heatmaps.

We just consider the case of locating one person?s one keypoint in an input image. This location can be represented spatially as a heatmap, and can be learned by a Fully Convolutional Networks (FCN) since it is simply a single-channel image. Therefore, if we want to predict all the keypoints in the image, we can set every keypoint type into a separate channel. We do same operation on connection heatmaps, and we only predict one connection in each channel.

Each pixel value in the keypoint heatmaps encodes the confidence that a nearest keypoint of a particular type occurs. And each pixel value in the connection heatmaps encodes whether this point is on the limb that two adjacent keypoints decide. We generate the ground truth keypoint heatmaps by putting Gaussian distributions with a standard deviation σ_k at all annotated keypoint positions. Left of Fig. 1 is an example of how we generate keypoint heatmaps.

We use improved PAFs to represent the connections between the two adjacent keypoints. The improved PAFs are used to encode the connection information between keypoints and extract the visual patterns of human skeleton. We follow IMHN that use an elliptical area to approximately represent the keypoints relationship. We generate the ground truth connection heatmaps by putting unnormalized elliptical Gaussian distributions with a standard deviation σ_p in all body part areas. Right of Fig. 1 is an example of how we generate connection heatmaps. The

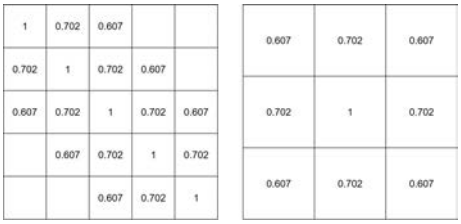

Fig. 1. Heatmap example. Left: PAFs, Right: Keypoint

standard deviations, σ_k and σ_p, control the spread of the Gaussian peaks and they should be set properly to balance the foreground pixels and background pixels. In our work, we set σ_k =9 and σ_p = 7 in the paper. The hyper-parameters $r0$ (keypoint radius) and $d0$ (body part width) determine the boundaries of the ground truth Gaussian peaks, truncating the unnormalized Gaussian distribution at a fixed value *thre*. It plays a role in our loss function.

3.2. Improved Part Affinity Fields

To address the problem that how to encode the connection between two keypoints, the Openpose approach presents a novel feature representation called part affinity fields that preserves both location and orientation information across the region of support of the limb. But Simple Pose [16] has found that it may bring vagueness or even conflicts to the information representation when all

the pixels within the approximate limb area (may include outliers of the limb) have the same ground truth value. Thus, they propose a new and more efficient connection encoding method called body part to evaluate the connection between keypoints.

The body part representation is more sensible and composite. Pixels will get higher confidence and score if they are near to the major axes of the body parts (Fig. 2). And it only needs the half dimensions of PAFs to encode the keypoint connection information, reducing the demand for model capacity. With the benefit, we combine the advantage of PAFs and body parts as Improved Part Affinity Fields (IPAFs) for our connection evaluation. To evaluate

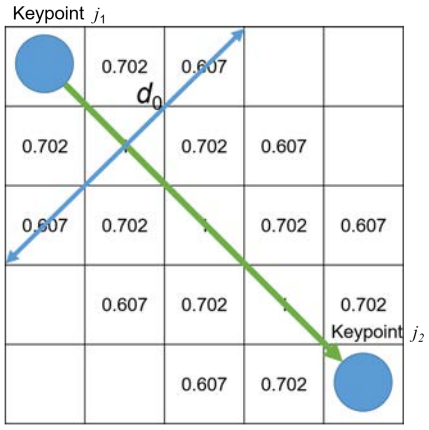

Fig. 2. Improved Part Affinity Fields

loss during training, we define the groundtruth connection, $L^*_{c,k}$, at an image point p as

$$L^*_{c,k}(p) = \begin{cases} v & p \in \text{limb}_{(c,k)} \\ 0 & \text{other} \end{cases} \quad \ldots\ldots \quad (1)$$

Here, the value of v depends on how close point p lies from the connection line between j_1 and j_2, it encodes the confidence of connection between j_1 and j_2, and valued between 0 and 1. And the set of points not on the limb is set to 0. The limb width d_0 is a distance valued by pixels. The groundtruth of improved part affinity field averages the affinity fields value of all persons in the image,

$$L^*_c(p) = \frac{1}{n_c(p)} \sum_k L^*_{c,k}(p) \quad \ldots\ldots\ldots \quad (2)$$

$n_c(p)$ represents the number of non-zero pixels at point p across all k persons, if limbs of different persons have overlapped, we calculate the average of all pixels' value.

In the testing phase, we first calculate the line integral over the corresponding IPAF, and then measure candidate part detections' relationship by calculated result. In our algorithm code, we calculate along the line segment connection of the candidate part locations. Since candidate connection could be formed by connecting the detected keypoints, we can measure the alignment of the predicted IPAF via the candidate connection. For instance, for two candidate part locations d_{j_1} and d_{j_2}, we sample the predicted improved part affinity field, L_c to measure the confidence in their line segment association.

$$E = \int_{u=0}^{u=1} L_c(p(u)) \bullet \frac{d_{j_2} - d_{j_1}}{\left\| d_{j_2} - d_{j_1} \right\|_2} du \quad \ldots\ldots \quad (3)$$

$p(u)$ interpolates the position of the two body parts d_{j_1} and d_{j_2},

$$p(u) = (1-u)d_{j_1} + ud_{j_2} \quad \ldots\ldots\ldots \quad (4)$$

In practice, we approximate the integral by sampling and summing uniformly-spaced values of u in our code.

3.3. Keypoint Assignment Algorithm

In order to predict multiple human pose in an image, we need all person keypoints and connections. We can obtain a discrete set of keypoint candidates by performing non-maximum suppression on the keypoint heatmaps. And similarly, we could get the candidate connections by scoring each candidate connection confidence using the line integral computation on the IPAFs. For each keypoint type, we may have several candidates for different person, because of multiple people or false detection in the image. And for each connection between keypoints, one keypoint may have several paired other keypoints. With the keypoints candidate and connections candidate, we could assemble a large set of possible limbs for pose generation. How to assign these possible limbs to corresponding person instance is a critical problem in assemble step. In this paper, we follow the greedy strategy in CMU-Pose and assemble the human skeletons by matching adjacent body parts independently.

Formally, we first obtain a set of keypoints candidates KJ from keypoint heatmaps for multiple people, defined as $K_j = \left\{ d_j^m : j \in \{1,\ldots,J\}, m \in \{1,\ldots,N_J\} \right\}$, with N_j is the number of candidates of keypoint j, and $d_j^m \in R^2$ is the location of the m-th detection candidate of keypoint j. We can use IPAFs make these part detection candidates knows which other part connected with. We define a variable $z^{mn}_{j_1 j_2} \in [0,1]$ to indicate the possibility that two detection candidates $d^m_{j_1}$ and $d^n_{j_2}$ are connected. So that we could find the connection of part detections that are in fact connected. After keypoint localization and part connection, we get a set of pairs of keypoints and then we could perform the assemble algorithm. So, this assignment problem can be transferred to find the optimal assignment for the set of all possible connection candidates, as the formula which we defined as $Z = \left\{ z^{mn}_{j_1 j_2} : j_1, j_2 \in \{1,\ldots,J\}, m \in \{1,\ldots,N_{j_1}\}, n \in \{1,\ldots,N_{j_2}\} \right\}$.

For instance, when we consider a single pair of parts j_1 and j_1 for the connection c, human skeleton has decided that each part has only one paired part, thus if one part is matched, there is no other part can be matched at same time. Bipartite graph matching problem has a property that in process of choosing a subset of edges we should ensure no two edges share a node. So, our assembling problem can be described as a maximum weight bipartite graph matching problem, the parts are the nodes and the

connection possibility are the weight of edges. Moreover, weight of each edge can be calculated by our IPAF aggregate. Consequently, our goal is choosing edges to make the whole graph with maximum weight.

$$\max_{Z_c} E_c = \max_{Z_c} \sum_{m\in K_{j_1}} \sum_{n\in K_{j_2}} E_{mn} \bullet z^{mn}_{j_1 j_2} \quad . \quad . \quad . \quad . \quad (5)$$

$$\forall m \in D_{j_1}, \sum_{n\in K_{j_2}} z^{mn}_{j_1 j_2} \leq 1 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (6)$$

$$\forall n \in D_{j_2}, \sum_{m\in K_{j_1}} z^{mn}_{j_1 j_2} \leq 1 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (7)$$

where E_c is the overall weight of the matching from limb type c, Z_c is the subset of Z for limb type c, E_{mn} is the connection possibility between parts $k^m_{j_1}$ and $k^n_{j_2}$. Eqs. (6) and (7) enforce no two edges share a node, and no two limbs of the same type (e.g., left forearm) share one part. We can obtain the optimal matching by perform the Hungarian algorithm.

When this problem generalizes from two between keypoints paired to assembling all human pose of multiple persons, the objective function Z becomes a K-dimensional matching problem and it is a NP Hard problem and relaxations exist too many to tackle with. In Openpose, they add two limitations to the optimization to simplify this problem. First, in order to save the computation cost of inference step, they choose a minimal number of edges enough to represent a person to obtain a spanning tree skeleton structure of human pose instead of using the fully connected graph. Second, they turn the full human pose assembling problem into a set of keypoints paired problems, which can be handled as bipartite matching subproblems and perform the Hungarian algorithm for several times to get the globally optimal solution.

With these two limitations, the optimization is decomposed simply as several bipartite graph matching problem:

$$\max_{Z} E = \sum_{c=1}^{C} \max_{Z_c} E_c \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (8)$$

While performing the Openpose keypoint assemble algorithm, we find a critical problem in inference step with the skeleton structure that Openpose defines. In the Openpose structure (Left of Fig.3), if any one of the connections 6,7,12,14,9,10 or the keypoints 3,9,12,6 has not been detected in case of occlusion or other reason, isolated part will be occurred which is critical for pose assemble. In our experiment, if isolated part appears, the original inference method that Openpose utilized will assign the isolated part to another nonexistent person consequently make the prediction accuracy worse. For tackling this problem, we proposed a new structure called Cycle Skeleton Structure (Right of Fig.3). We add connections 20, 21, 22 to make sure that if only one keypoint or limb has not been found in one person, there is another connection ensuring the right assignment. The three edges we add make all keypoints in a loop cycle, so that we call this structure Cycle Skeleton Structure. And we also delete the unnecessary connections in the region of head for making up for the efficiency loss of added connections cause.

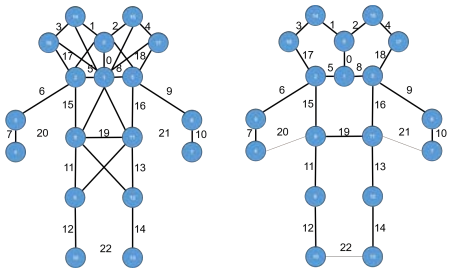

Fig. 3. Keypoint Assignment. Left: PAFs, Right: Ours

3.4. Network structure

We use IMHN as our main network architecture for our idea. It takes an image of any shape as the input and outputs multi-scale keypoint and body part heatmaps of all persons (if any) in the scene simultaneously by Hourglass module [16]. Before feeding into the stacked Hourglass modules, the original input is down-sampled to quarter size by some convolutional layers and max-pooling layers. Output has 43 channels, the keypoint heatmaps have 18 channels, corresponding to the keypoint types (we follow Openpose to add neck keypoint to annotation). And IPAFs heatmaps have 23 channels. The rest of 2 channels is background channel and foreground channel to balance the sample distribution. The network structure is shown in Fig. 4.

Fig. 4. Network architecture

4. EXPERIMENTS

4.1. Dataset

The COCO [14] dataset contains over 200, 000 images and 250, 000 person instances labeled with 17 keypoints. We train our model on COCO train2017 dataset, including 57K images and 150K person instances. We evaluate our approach on the val2017 set, containing 5000 images.

4.2. Evaluation metric

The standard evaluation metric is based on Object Keypoint Similarity (OKS):

$$OKS = \frac{\sum_i \exp(d_i{}^2/2s^2k_i{}^2)\delta(v_i > 0)}{\sum_i \delta(v_i > 0)} \quad . \quad . \quad . \quad . \quad (9)$$

Here d_i is the Euclidean distance between the detected keypoint and the corresponding ground truth, v_i is the visibility flag of the ground truth, $v_i > 0$ means this keypoint is visible, s is the object scale, and k_i is a per-keypoint constant that controls falloff. As the definition

of *OKS*,*OKS* is utilized for evaluating two keypoints how similar to each other. Following the COCO dataset evaluation API, we report standard average precision(AP) and average recall(AR) scores 1: AP50(AP at *OKS* = 0.50), AP75 (AP at *OKS*= 0.75), mAP (the mean of AP scores at 10 positions, *OKS* = 0.50, 0.55, . . ., 0.90, 0.95; and mAR at *OKS* = 0.50, 0.55, . . ., 0.90, 0.95.

4.3. Training detail

For our computation limitation, we have built the model with pytorch and train it on art of the COCO2017 keypoint training dataset for 25 epochs. Each epoch has 1000 images. The relationship between epoch and loss is indicated in Fig.5.

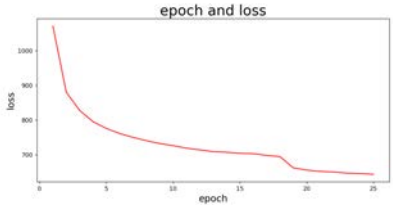

Fig. 5. Loss in training

4.4. Results on the COCO2017 validation set

We have tested our approach on a subset of COCO2017 validation set. We choose the subset images at the principle that Openpose performs badly on it. We report the results of our method and other state-of–the-art methods in Table.1.

Table 1. Results on the subset of COCO2017 validation set.

Method	mAP(%)	AP50(%)	AP75(%)	mAR(%)
Openpose	60.6	80.2	64.7	63.0
IMHN	62.8	80.6	72.5	64.5
Ours	68.9	85.2	78.0	70.9

We perform our improvement with visualization form (Fig. 6). We make a handcraft occlusion mask on the person in the image in order to let one connection cannot be detected. In left box of Fig. 6, when we perform the Openpose approach, it does not assemble the left leg into the whole pose so that two persons will be detected as false result (each red box represent a detected person). In right box of Fig. 6, we make a mask that cover the upper part of the body. In this situation, the Openpose approach could not connect the isolated parts together and make false prediction (it has detected 7 persons despite only 2 persons exist in the image). In contrary, in spite of not detecting occluded connections, our approach handles the isolated part problem well and gives out the right result. Moreover, in our speed test, we got 7.5 FPS in inference process better than 8.1 FPS of IMHN in one 1070 GPU and i7-7820HK CPU. We have average 6% speed up in inference process.

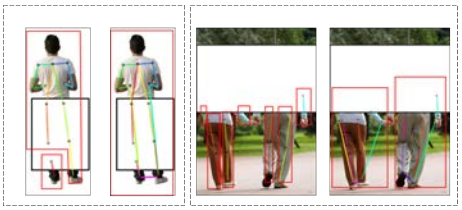

Fig. 6. Loss in training

5. CONCLUSIONS

In this paper, we develop a bottom-up approach for multi-person pose estimation. We provide some insights into valuable design choices: (1) we develop a new method called IPAFs to evaluate connection of two adjacent keypoints, (2) we use a powerful network to generate multi scale heatmaps, and (3) we adjust the skeleton graph to make pose assemble phase robust. In the future, we will try more method to improve our accuracy and speed up our approach.

Acknowledgements

This work was supported by the Opening Fund of Key Laboratory of Geological Survey and Evaluation of Ministry of Education (Grant No. GLAB2020 ZR06) and the Fundamental Research Funds for the Central Universities, the Joint Foundation of China Aerospace Science and Industry for Equipment Pre Research 2020, and the National Natural Science Foundation of China under contracts 61603357.

References:

[1] Sun K, Xiao B, Liu D, Wang J, “Deep High-Resolution Representation Learning for Human Pose Estimation”, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5693–5703, 2019.

[2] Xiao B, Wu H, Wei Y, “Deep High-Resolution Representation Learning for Human Pose Estimation”, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5693–5703, 2019.

[3] Papandreou G, Zhu T, Chen L C, Gidaris S, Tompson J, Murphy K, “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, 2018.

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

[5] Newell A, Huang Z, Deng J, “Associative embedding: End-to-end learning for joint detection and grouping”, in Proceedings of Advances in Neural Information Processing Systems, pp. 2277–2287, 2017.

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

[7] Li J, Su W, Wang Z, “Simple Pose: Rethinking and Improving a Bottom-up Approach for Multi-Person Pose Estimation”, in Proceedings of the National Conference on Artificial Intelligence (AAAI), pp. 11354–11361, 2020.

[8] Newell, A.; Yang, K.; Deng, J, “Stacked hoursglass networks for human pose estimation”, in Proceedings of European Conference on Computer Vision (ECCV), pp. 483–499, 2016.

[9] Chen Y, Wang Z, Peng Y, Zhang Z, Yu G, Sun Ji, “Cascaded pyramid network for multi-person pose estimation”, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7103–7112, 2018.

[10] Fang H S, Xie S, Tai Y W, Lu C, “Rmpe: Regional multi-person pose estimation”, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2334–2343, 2017.

[11] Papandreou G, Zhu T, Kanazawa N, Toshev A, Tompson T, Bregler C, et al, “Towards accurate multi-person pose estimation in the wild”, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4903–4911, 2017.

[12] Li J, Wang C, Zhu H, Mao Y, Fang H S, Lu C, “Crowdpose: Efficient crowded scenes pose estimation and a new benchmark”, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10863–10872, 2019.

[13] Nie X, Feng J, Zhang J, Yan S, “Single-stage multi-person pose machines”, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6951–6960, 2019.

[14] Lin T Y, Maire M, Belongie S, Bourdev L, Hays J, Perona P, et al, “Microsoft coco: Common objects in context”, in Proceedings of European Conference on Computer Vision(ECCV), pp. 740–755, 2014.