Keypoint Context Aggregation for Human Pose Estimation

Wenzhu Wu[1,2], Weining Wang[1,2], Longteng Guo[1,2], and Jing Liu[1,2](✉)

[1] National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, China

[2] School of Artificial Intelligence, University of Chinese Academy of Sciences,
Beijing, China

wuwenzhu2019@ia.ac.cn, {weining.wang,longteng.guo,jing.liu}@nlpr.ia.ac.cn

Abstract. Human pose estimation has drawn much attention recently, but it remains challenging due to the deformation of human joints, the occlusion between limbs, etc. And more discriminative feature representations will bring more accurate prediction results. In this paper, we explore the importance of aggregating keypoint contextual information to strengthen the feature map representations in human pose estimation. Motivated by the fact that each keypoint is characterized by its relative contextual keypoints, we devise a simple yet effective approach, namely Keypoint Context Aggregation Module, that aggregates informative keypoint contexts for better keypoint localization. Specifically, first we obtain a rough localization result, which can be considered as soft keypoint areas. Based on these soft areas, keypoint contexts are purposefully aggregated for feature representation strengthening. Experiments show that the proposed Keypoint Context Aggregation Module can be used in various backbones to boost the performance and our best model achieves a state-of-the-art of 75.8% AP on MSCOCO test-dev split.

Keywords: Human pose estimation · Keypoint context · Feature augmentation

1 Introduction

Multi-person human pose estimation aims at recognizing and localizing the anatomical keypoints of all persons in a given image. As one of the most fundamental tasks in computer vision, it serves as a key component for many other vision applications, including human action recognition, human-computer interaction, virtual or augmented reality, etc. Despite the noticeable improvements achieved in this area by advanced deep learning techniques [7,8,17,18,20,22,25,30], pose estimation still remains extremely challenging. It is still difficult to locate the keypoint coordinates precisely, due to the variation of clothing, the occlusion between the limbs and the deformation of human joints under different poses.

To address the above problems, full utilization of contextual information has been widely concerned. A line of previous works focus on exploring multi-scale

Y. Peng et al. (Eds.): ICIG 2021, LNCS 12889, pp. 386–396, 2021.
https://doi.org/10.1007/978-3-030-87358-5_31

contextual information to enhance the performance of keypoint localization. [15, 20] aggregated multi-scale contexts that generated from well-designed interactions among feature maps of different levels and resolutions. Another line of works are from the perspective of promoting the information communication within a local scope. [6] designed a hierarchical visual attention scheme to zoom in a smaller body part, which generated a specific attention map for each body joint. [1] proposed a novel network - RSN, which aims to learn delicate local representations by efficient intra-level feature fusion.

Although the existing works exploit contextual information from different perspectives, none of them explore keypoint areas as context. However, keypoint context is crucial for precise keypoint localization. Each position on the feature map contains different keypoint information, for example, a specific pixel may contain more characteristics of the left shoulder than the left ankle. Therefore, more characterized keypoint context information aggregation helps to make more accurate localization for prediction. Motivated by this, in this paper, we propose a Keypoint Context Aggregation Module (KCAM) to leverage the relationships between pixels and its contextual keypoints.

For each pixel on the image, KCAM can effectively learn the relationships between it and all human keypoints, thus we can aggregate the keypoint representations purposefully for the current pixel according to their relationships. The augmented feature can lead to more accurate localization. The whole scheme for KCAM can be summarized as follows. Firstly, we obtain a rough keypoint localization result, which can be considered as soft keypoints. Secondly, based on the coarse keypoint localization, we acquire the informative keypoint representations. Thirdly, we compute the relationship between each pixel and all keypoint representations. For the last step, for each pixel, we aggregate keypoint representations purposefully according to the calculated relationships, thus obtaining the augmented feature representation. The augmented feature is then used for final precise keypoint localization. We evaluate the proposed method on the benchmark dataset MSCOCO [16], and experimental results demonstrate the effectiveness of KCAM.

In summary, our main contributions are three-fold as follows:

- We propose a Keypoint Context Aggregation Module for human pose estimation, which can effectively aggregate representative and informative keypoint contextual information for reasonable feature augmentation and conduce more accurate keypoint localization results.
- Keypoint Context Aggregation Module can serve as a model-agnostic refinement method, which can be easily applied to the existing pose estimation methods.
- Our method outperforms state-of-the-art methods on the challenging benchmark MSCOCO dataset for human pose estimation.

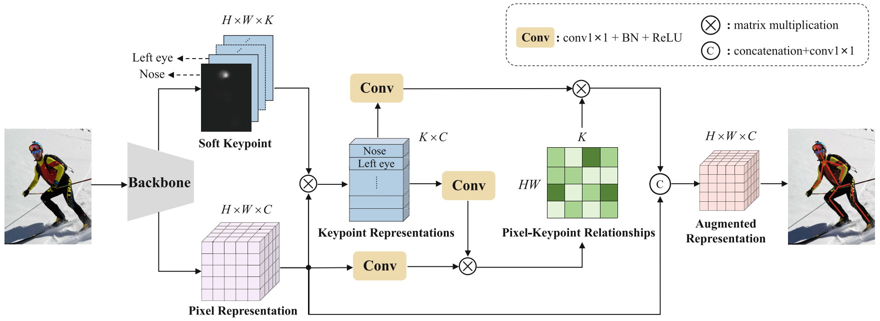

Fig. 1. Illustration of the pipeline of our proposed Keypoint Context Aggregation Module. (a) Acquire rough localization results for soft keypoint areas with intermediate supervision. (b) Obtain the informative keypoint representations. (c) Compute keypoint contextual representations and acquire the aggregated representation. (d) Concatenate original feature representation and the aggregated representation, followed by a conv 1 × 1 layer to get the final augmented feature representation.

2 Related Work

Multi-person Pose Estimation. For multi-person pose estimation, it can be classified into top-down and bottom-up methods. Top-down methods [4,9,11,24–26,29] construct human body poses by detecting the people first and then apply single-person pose estimators to predict the keypoints for each person. Different from top-down methods, bottom-up methods [2,5,10,13,14,19,21,23] detect all the body joints in one image first and then group them into individual poses. In the top-down pipeline, the number of people in the input image will directly affect the computing time. The computing speed for bottom-up methods is usually faster than top-down methods since they do not need to detect the pose for each person separately.

Relational Context. Previous works always explore the relationships among different keypoints. Zhang et al. [31] build a pose graph directly on keypoint heatmaps and use Graph Neural Network for modeling, which only considers the relationship between heatmap weights at the same location, while et al. [27] build pose graph considering the visual features at the position of corresponding keypoints. The above methods are dedicated to strengthening the relationships between keypoints. However, our approach is quite different, that we explore the relationships between feature maps and keypoints.

Coarse-to-Fine Pose Estimation. Various coarse-to-fine pose estimation schemes have been developed to gradually refine the result heatmaps from coarse to fine. Carreira et al. [3] refine pose estimation by predicting error feedback at each iteration, [1,2,20] design a cascaded architecture for mining multi-stage prediction. Our approach in some sense can also be regarded as a coarse-to-fine scheme. The difference lies in that we use the coarse segmentation map for generating a contextual representation instead of directly used as an extra representation.

3 Proposed Method

The overall pipeline of our method is illustrated in Fig. 1. First of all, the input image is sent to the backbone network. Then, taking the backbone output as input, the proposed Keypoint Context Aggregation Module can be summarized as: Firstly, heatmap regression network is applied to acquire rough localization results, which are considered as soft keypoint areas. Secondly, based on the keypoint areas, we obtain the informative keypoint representations. Thirdly, we compute the relationships between each pixel on the feature map and all keypoint representations. Then, we aggregate the feature representations with keypoint context through the computed relationships. Finally, with feature fusion, we acquire the augmented features for precise localization.

3.1 Keypoint Representations

Considering that we will leverage keypoint areas information to augment feature representations, it is essential to roughly locate the keypoint first. We predict K keypoint areas $\{G_1, ..., G_K, G_k \in \mathbb{R}^{H \times W}\}$ from the output feature map $F \in \mathbb{R}^{C \times H \times W}$ of the input image I. Each keypoint area can be described as a 2D heatmap whose per-pixel value g_{ik} indicates the probability of the k-th ($k = 1, 2, ...K$) keypoint's presence at this location i. In order to get more accurate location, we use intermediate supervision here and take the keypoints Guassian maps as ground-truth.

After obtaining the soft keypoint areas, we acquire the keypoint representations $R \in \mathbb{R}^{K \times C}$ through the following formulation:

$$R_k = \sum_{i=1}^{HW} g_{ik} f_i \tag{1}$$

where g_{ik} is the normalized degree for i-th pixel belonging to the k-th keypoint and $f_i \in \mathbb{R}^{1 \times C}$ indicates the representation of this pixel. And $R_k \in \mathbb{R}^{1 \times C}$ is the representation for k-th keypoint.

3.2 Pixel-Keypoint Relationships

In order to model rich contextual relationships, we encode contextual keypoint information into pixel features, thus enhancing feature representations capability. First, we compute the relationships between the feature of each pixel and keypoint representations. Given the original visual feature $F \in \mathbb{R}^{C \times H \times W}$, keypoint representations $R \in \mathbb{R}^{K \times C}$, we firstly feed it into a convolution layer to generate two new features as $F' \in \mathbb{R}^{C\times H\times W}, R' \in \mathbb{R}^{K\times C}$, respectively. After reshaping F to $\mathbb{R}^{C\times N}$, where $N = W \times H$ we apply a matrix multiplication between the transpose of F' and R', and apply a softmax layer to calculate the relationships $S \in \mathbb{R}^{N\times K}$ between pixels and keypoint representations:

$$S_{ij} = \frac{exp(F'_i \times R'_j)}{\sum_{j=1}^{K} exp(F'_i \times R'_j)} \tag{2}$$

where S_{ij} indicates j-th keypoint's impact on i-th pixel. The greater the attention weight is, the more the pixel is related with the corresponding keypoint representation.

3.3 Augmented Representation

Meanwhile, keypoint representations $R \in \mathbb{R}^{K\times C}$ is fed into a convolution layer to generate a new feature $R'' \in \mathbb{R}^{K\times C}$. Based on the above attention map, we compute the aggregated representation $E \in \mathbb{R}^{C\times H\times W}$ as the weighted sum of projected keypoint representations:

$$E_i = \sum_{j=1}^{K} S_{ij} R''_j \tag{3}$$

Then, We combine F and E through concatenation operation, and feed it to a conv 1×1 layer to get the final feature representation $A \in \mathbb{R}^{C\times H\times W}$. The resulting features are used for final keypoints localization.

3.4 Overall Loss Function

The overall loss is composed of two keypoints heatmap losses: one for intermediate supervision result, the other for the final output. The loss function can be described as:

$$L = \alpha L_{inter} + L_{output} \tag{4}$$

where the hyperparameter α is set to 0.2. Both L_{inter} and L_{output} are computed as:

$$L = \frac{1}{N} \sum_{n=1}^{N} \sum_{x,y} ||P_n(x,y) - G_n(x,y)||_2 \tag{5}$$

where $P_n(x,y)$ and $G_n(x,y)$ represent the predicted and the ground-truth confidence maps at the pixel location (x,y) for the n-th keypoint, respectively.

4 Experiments

4.1 Dataset

Our experiments are conducted on human keypoint detection task of the large-scale benchmark MSCOCO dataset [16]. The dataset contains over 200K images and 250K person instances labeled with 17 keypoints. Following the common practice [1], we train our model on MSCOCO trainval dataset (includes 57K images and 150K person instances) and validate on MSCOCO minival dataset (includes 5000 images). We report our main results on the test-challenge set (20K images). Unless specified, we only make use of the human keypoint annotations without bounding-boxes. The performance is computed with Average Precision (AP) based on Object Keypoint Similarity (OKS).

4.2 Implementaion Details

We use two different input sizes (256 × 192, 384 × 288) in our experiments. We initialize the backbones using the model pre-trained on ImageNet. Meanwhile, we initialize the Keypoint Context Aggregation Module randomly. The data augmentation includes random rotation ([$-40°$, $40°$]), random scale ([0.5, 1.5]), and flipping. Following [28], half-body data augmentation is also involved. All models are trained using 4-GPU machines, with a batch size of 128 images. Batch normalization is used in our network. We use the Adam optimizer [12]. The base learning rate is set as 5e-4, and is dropped to 5e-5 and 5e-6 at the 170th and 200th epochs, respectively. The training process is terminated within 210 epochs.

We use the same person detectors provided by SimpleBaseline [29] for both validation and test-dev set. Following the same techniques used in [4], we also predict the pose of the corresponding flipped image and average the heatmaps to get the final prediction; a quarter offset in the direction from the highest response to the second highest response is used to obtain the final location of the keypoints.

4.3 Ablation Studies

We conduct the empirical analysis on MSCOCO validation set. Unless specified, all the ablation experiments are based on the backbone of Resnet-50 with the input size of 256 × 192.

Table 1. Influence of the soft keypoint supervision scheme. We can find that the soft keypoint supervision scheme is important for the performance.

Method	AP	AP_{50}	AP_{75}	AP_M	AP_L
w/o supervision	70.19	89.59	78.10	66.27	76.99
w/ supervision	**71.87**	**89.77**	**79.89**	**68.26**	**78.45**

Keypoint Region Supervision. We study the influence of the supervision for keypoint areas. We modify our approach by removing the supervision on the soft keypoint areas. We keep all the other settings unchanged and report the results in the Table 1. From the result, we can see that the performance decreased by a large margin without the intermediate supervision. Thus, it can be inferred that the intermediate supervision for the keypoints is crucial for KCAM. The reason is that more accurate localization of the keypoint areas will lead to more informative and representative keypoint representations.

Table 2. Influence of the different feature fusion operations. We can find that concatenation operation is the most effective.

Method	AP	AP_{50}	AP_{75}	AP_M	AP_L
Concatenation	**71.87**	89.77	**79.89**	**68.26**	78.45
Element-wise addition	71.77	89.82	79.41	68.11	**78.50**
w/o fusion	70.94	**89.92**	79.37	67.22	77.61

Feature Fusion. We explore the fusion operation for aggregating the original visual features and the augmented features. Here we discuss the influence of several different ways to fuse these two features. We finally choose the fusion operation which achieves the better performance on AP metric, for AP metric is an overall metric for measuring human pose estimation model performance and is much more important than other metrics. As shown in Table 2, feature concatenation achieves an AP of 71.87, which slightly surpasses element-wise addition operation by an AP of 0.1. Additionally, we study the result of only using the augmented features for final prediction, whose performance decreased by an AP of 0.93. The above results show that the feature fusion is necessary and concatenation operation leads to better performance.

Visualization Analysis. As shown in Fig. 2, we visualize the relationships between the chosen pixel (indicated by stars) and contextual keypoint areas (indicated by circles). Red, blue and other colors denote 1, 0, and the values between them, respectively. The closer to red, the more related to the current keypoint. So the left one shows us, the center point of the person is more relevant to the shoulder and hip, which is spatially closer to it. From the right one we can conclude that, the point on the left calf is more related to the keypoints on limbs. The above results further show that our proposed method effectively aggregates the keypoints context. Incorporating with the keypoint context helps to distinguish the keypoint type for the current pixel. Figure 3 illustrates some results generated using our method. In multi-person scenes, keypoints occluded by clothes or other limbs can also be accurately located.

4.4 Comparison with State-of-the-Art

We compare our method with top-performers including G-RMI [22], CPN [4], HigherHRNet [5], SimpleBaseline [29], and HRNet [25]. Table 3 shows the accuracy results of these state-of-the-art methods and KCAM on the MSCOCO test-dev set. In this test, we use the person detection results from [29]. We have

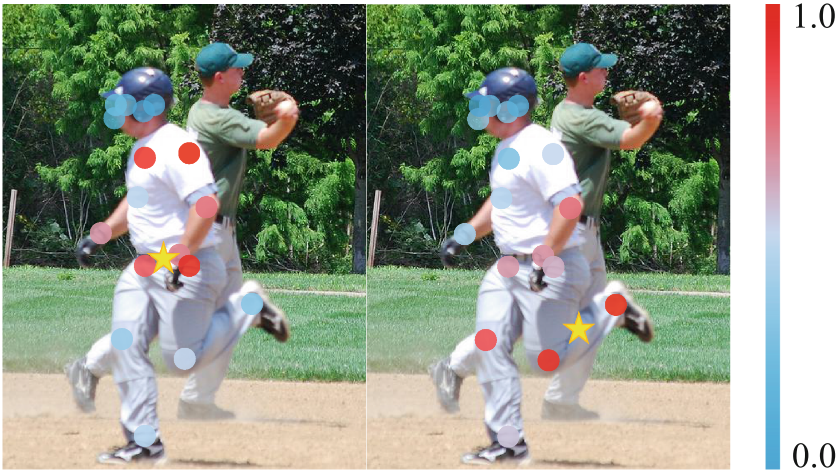

Fig. 2. Visualization of the relationships between the chosen pixel and contextual keypoint areas. The circles indicate the keypoint areas, while the star indicates the chosen pixel. Red, blue and other colors denote 1, 0, and the values between them, respectively. (Color figure online)

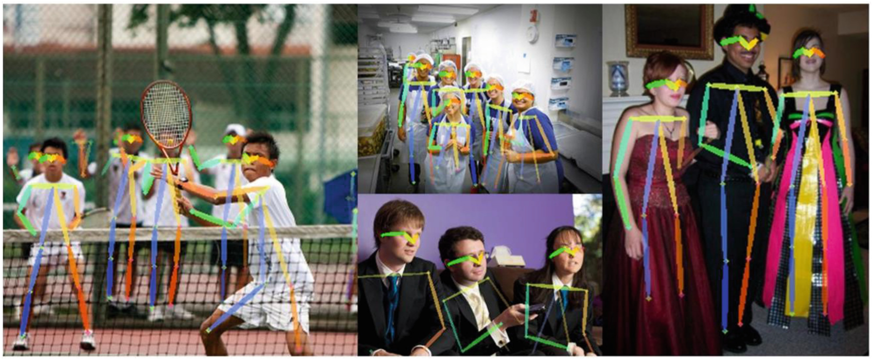

Fig. 3. Visualization of the results predicted by our models.

observed that KCAM with HRNet-W48 at the input size of 384 × 288 achieves the best accuracy. Specifically, compared with the best competitor (HRNet-W48 with the same input size), KCAM further improves AP by 0.3. The result again illustrates the effectiveness of our method.

Table 3. Comparisons on MSCOCO test-dev dataset. (+) indicates models ensembled.

Method	Backbone	Input size	AP	AP_{50}	AP_{75}	AP_M	AP_L
CMU-Pose [2]	–	–	61.8	84.9	67.5	57.1	68.2
Mask-RCNN [7]	ResNet-50-FPN	-	63.1	87.3	68.7	57.8	71.4
G-RMI [22]	ResNet-101	353 × 257	64.9	85.5	71.3	62.3	70.0
AE [19]	–	–	65.5	86.8	72.3	60.6	72.6
PersonLab [21]	–	–	68.7	89.0	75.4	64.1	75.5
HigherHRNet [5]	HRNet-W48	640 × 640	70.5	89.3	77.2	66.6	75.8
CPN [4]	ResNet-Inception	384 × 288	72.1	91.4	80.0	68.7	77.2
CPN$^+$ [4]	ResNet-Inception	384 × 288	73.0	91.7	80.9	69.5	78.1
SimpleBaseline [29]	ResNet-152	384 × 288	73.7	91.9	81.1	70.3	80.0
HRNet-W48 [25]	HRNet-W48	384 × 288	75.5	92.5	83.3	71.9	81.5
Ours	HRNet-W48	384 × 288	**75.8**	**92.7**	**83.6**	**72.3**	**81.8**

5 Conclusion

In this work, we have proposed a Keypoint Context Aggregation Module for human pose estimation, which can effectively aggregate representative and informative keypoint contextual information for reasonable feature augmentation and conduce more accurate keypoint localization results. We empirically show that our approach brings consistent improvements on MSCOCO benchmark.

References

1. Cai, Y., et al.: Learning delicate local representations for multi-person pose estimation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12348, pp. 455–472. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58580-8_27
2. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. In: CVPR, pp. 7291–7299 (2017)
3. Carreira, J., Agrawal, P., Fragkiadaki, K., Malik, J.: Human pose estimation with iterative error feedback. In: CVPR, pp. 4733–4742 (2016)
4. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: CVPR, pp. 7103–7112 (2018)
5. Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T.S., Zhang, L.: Higherhrnet: scale-aware representation learning for bottom-up human pose estimation. In: CVPR, pp. 5386–5395 (2020)
6. Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A.L., Wang, X.: Multi-context attention for human pose estimation. In: CVPR, pp. 1831–1840 (2017)
7. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: ICCV, pp. 2961–2969 (2017)
8. Huang, J., Zhu, Z., Guo, F., Huang, G.: The devil is in the details: delving into unbiased data processing for human pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5700–5709 (2020)

9. Iqbal, U., Gall, J.: Multi-person pose estimation with local joint-to-person associations. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9914, pp. 627–642. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48881-3_44
10. Jin, S., et al.: Differentiable hierarchical graph grouping for multi-person pose estimation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12352, pp. 718–734. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58571-6_42
11. Jin, S., et al.: Whole-body human pose estimation in the wild. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12354, pp. 196–214. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58545-7_12
12. Kingma, D.P., Ba, J.L.: Adam: a method for stochastic gradient descent. In: ICLR, pp. 1–15 (2015)
13. Kocabas, M., Karagoz, S., Akbas, E.: Multiposenet: fast multi-person pose estimation using pose residual network. In: ECCV, pp. 417–433 (2018)
14. Kreiss, S., Bertoni, L., Alahi, A.: Pifpaf: composite fields for human pose estimation. In: CVPR, pp. 11977–11986 (2019)
15. Li, W., et al.: Rethinking on multi-stage networks for human pose estimation. arXiv preprint arXiv:1901.00148 (2019)
16. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
17. Luvizon, D.C., Tabia, H., Picard, D.: Human pose regression by combining indirect part detection and contextual information. Comput. Graph. **85**, 15–22 (2019)
18. Moon, G., Chang, J.Y., Lee, K.M.: Posefix: model-agnostic general human pose refinement network. In: CVPR, pp. 7773–7781 (2019)
19. Newell, A., Huang, Z., Deng, J.: Associative embedding: end-to-end learning for joint detection and grouping. In: NeurIPS, pp. 2277–2287 (2017)
20. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 483–499. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_29
21. 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: ECCV, pp. 269–286 (2018)
22. Papandreou, G., et al.: Towards accurate multi-person pose estimation in the wild. In: CVPR, pp. 4903–4911 (2017)
23. Pishchulin, L., et al.: Deepcut: joint subset partition and labeling for multi person pose estimation. In: CVPR, pp. 4929–4937 (2016)
24. Qiu, L., et al.: Peeking into occluded joints: a novel framework for crowd pose estimation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12364, pp. 488–504. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58529-7_29
25. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: CVPR, pp. 5693–5703 (2019)
26. Umer, R., Doering, A., Leibe, B., Gall, J.: Self-supervised keypoint correspondences for multi-person pose estimation and tracking in videos. arXiv preprint arXiv:2004.12652 (2020)
27. Wang, J., Long, X., Gao, Y., Ding, E., Wen, S.: Graph-PCNN: two stage human pose estimation with graph pose refinement. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12356, pp. 492–508. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58621-8_29

28. Wang, Z., et al.: Mscoco keypoints challenge 2018. In: Joint Recognition Challenge Workshop at ECCV (2018)
29. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: ECCV, pp. 466–481 (2018)
30. Zhang, F., Zhu, X., Ye, M.: Fast human pose estimation. In: CVPR, pp. 3517–3526 (2019)
31. Zhang, H., et al.: Human pose estimation with spatial contextual information. arXiv preprint arXiv:1901.01760 (2019)