Exploiting Offset-guided Network for Pose Estimation and Tracking

Rui Zhang*[1,2], Zheng Zhu*[3], Peng Li[1], Rui Wu[4], Chaoxu Guo*[3], Guan Huang[4], Hailun Xia*[1,2]
[1] Bejing Key Laboratory of Network System Architecture and Convergence,
School of Information and Communication Engineering,
Beijing University of Posts and Telecommunications, Beijing.
[2] Beijing Laboratory of Advanced Information Networks, Beijing.
[3] Institute of Automation, Chinese Academy of Sciences, Beijing.
[4] Horizon Robotics, Beijing.
zhengzhu@ieee.org

Abstract

Human pose estimation has witnessed a significant advance thanks to the development of deep learning. Recent human pose estimation approaches tend to directly predict the location heatmaps, which causes quantization errors and inevitably deteriorates the performance within the reduced network output. Aim at solving it, we revisit the heatmap-offset aggregation method and propose the Offset-guided Network (OGN) with an intuitive but effective fusion strategy for both two-stages pose estimation and Mask R-CNN. For two-stages pose estimation, a greedy box generation strategy is also proposed to keep more necessary candidates while performing person detection. For mask R-CNN, ratio-consistent is adopted to improve the generalization ability of the network. State-of-the-art results on COCO and PoseTrack dataset verify the effectiveness of our offset-guided pose estimation and tracking.

1. Introduction

Human pose estimation in images and articulated pose tracking in videos are of significance for visual understanding task [32, 12]. Research community has witnessed a significant advance from single person [3, 10, 27, 26, 28, 21, 31] to multi-person pose estimation [23, 15, 4, 22, 5], from static images pose estimation [24, 12] to articulated tracking in videos [16, 14, 8, 17, 34, 11, 30, 29]. However, there are still challenging pose estimation problems in complex environments, such as occlusion, intense light and rare poses [2, 18, 25]. Furthermore, articulated tracking encounters new challenges in unconstrained videos such as camera motion, blur and view variants [1, 33, 35].

*The first two authors contributed equally to this work.

Previous pose estimation systems address single pre-located person, which exploit pictorial structures model[3, 10] and deep convolutional neural network [27, 26, 28, 21, 31]. Motivated by practical applications in video surveillance, human-computer interaction and action recognition, researchers now focus on the multi-person pose estimation in unconstrained environments. Multi-person pose estimation can be categorized into bottom-up [23, 15, 4] and top-down approaches [22, 5, 12, 29], where the latter becomes dominant participants in COCO benchmarks [18]. Top-down approaches can be divided into two-stages based methods and unified framework. Two-stages methods [22, 5, 29] firstly detect and crop persons from the image, then perform the single person pose estimation in the cropped person patches. Representative work of unified framework methods is Mask R-CNN [12], which extracts the human bounding box and predicts keypoints from the corresponding feature maps simultaneously.

While there has been a significant advance in pose estimation, quantization errors still exist in most of the modern networks. Although Google [22] proposes to simultaneously classify the heatmaps and regress the offset filed, recent human pose estimation approaches [12, 5, 29] tend to directly predict the location heatmaps. Because of the quantization effect between input and output, performance is inevitably deteriorated within the reduced network output. While both deconv and offset can reduce quantization errors, offset is more significant for resources-restricted applications due to its efficiency. In this paper, we revisit the heatmap-offset aggregation method and propose the Offset-guided Network (OGN) for both two-stages pose estimation and unified Mask R-CNN framework. We extend modern frameworks by adding a branch for offset prediction in parallel with the existing branch. Meanwhile, an intuitive but effective fusion is adopted to obtain the final results, and we

propose a greedy box generation strategy to keep more necessary candidates. The OGN aims at improving precision for all sizes output especially low resolution. Our network can output keypoints location in continuous space which reduces the quantization error.

In experiments, the offset-guided two-stages pose estimation approach reaches mAP of 74.0 on COCO test-dev set, yielding 14% relative gain compared with [22]. On PoseTrack dataset, we achieve 67.7 MOTA using two-stages pose input without optical flow, which is the new state-of-the-art results in this task.

The main contributions can be described as follows:

(1) Heatmap-offset aggregation method is revisited and we propose the OGN for both two-stages pose estimation and Mask R-CNN. An intuitive but effective fusion strategy is adopted to obtain the final results by merging two branches.

(2) As a novel alternative to NMS, a greedy box generation strategy is adopted to keep more necessary candidates for offset-guided two-stages pose estimation.

(3) In experiments, the offset-guided two-stages pose estimation approach reaches mAP of 74.0 on COCO test-dev set with a single model, yielding 14% relative gain compared to [22]. Furthermore, we achieve 67.7 MOTA on PoseTrack dataset without optical flow, which is the new state-of-the-art results in this task.

2. Related Works

2.1. Single person pose estimation

Single person pose estimation is a task that predicts the pose of a single person in an image. Conventional methods [3, 10] exploit pictorial structure model which expresses the human body as a tree-structured graphical model. [3] claims that the right selection of components for both appearance and spatial modeling is crucial. The Deformable Part Model (DPM) [10] adopts HOG feature to implement this idea. Recently, this task has been advanced rapidly for the development of deep convolution neural networks. [27] firstly tries to utilize CNN and they prefer to directly regress coordinates of body parts. More recently, researches on this task choose to regress some heat maps, which each stands for a body part. [26] is the first work which solves the problem by using CNN and graphical models to predict heat maps of each body part. With the continuous work of many researchers, some novel architectures like CPM [28], Stacked Hourglass [21] and PRMs [31] are used to achieve state-of-the-art results.

2.2. Multi-person pose estimation

Motivated by practical applications, researchers now focus on multi-person in unconstrained environments. Multi-person pose estimation can be categorized into bottom-up and top-down approaches where the latter becomes dominant participants in COCO benchmarks [18].

bottom-up Bottom-up architecture based methods first detect body parts and then associate corresponding body parts with specific human instances. The typical methods are DeepCut [23] and DeeperCut [15]. The former adopts an integer linear programming based method and the later improves DeepCut via utilizing image-conditioned pairwise terms. [4] predicts heatmaps of body parts and a set of 2D vector fields of part affinities and parses them by greedy inference to generate the final results.

Top-down Top-down approaches can be divided into two-stages based methods and unified framework. Two-stages methods [24, 22, 5, 29] first detect and crop persons from an image, then perform single person pose estimation in the cropped person patches. [24] follows this two-step framework by using pictorial structure models based method. [22] combines classification and regression tasks which respectively predicts the offset vector and location heatmap of each body part. [5] proposes a cascaded pyramid network containing global pyramid network and pyramid refined network which aims for online hard key points mining. Representative work of unified framework methods is Mask R-CNN [12] that builds an end-to-end framework and yields an impressive performance.

3. Overview of offset-guided network

For pose estimation, it is noticed that the precision of keypoints localization is limited by the size of network output. During the downsampling process, there exists a quantization error. OGN is utilized to address this problem. We verify the effectiveness of OGN for two-stages pose estimation (shown in Figure 1) and extended Mask R-CNN framework (Figure 2). Following [22], the regions of interest (ROIs) detected and cropped by person detector are fed to the pose estimator, where the offset regression branch guides heatmap classification branch to refine the pose location. Differently, two *deconv* layers [29] are used to enlarge the heatmaps by four times and an intuitive but effective fusion is adopted to obtain the final results. Meanwhile, in extended Mask R-CNN as shown in Figure 2, the ROIs from RPN are firstly extended to a fixed ratio and then ROI-Align is used to extract the feature in each extended ROIs. Finally, a score map and an offset map are predicted and fused to obtain the final location of keypoints.

3.1. Offset-guided two-stages Pose Estimation

We first address the OGN for two-stages pose estimation framework. For the first stage, the results of the person detector are crucial for subsequent pose estimator. However,

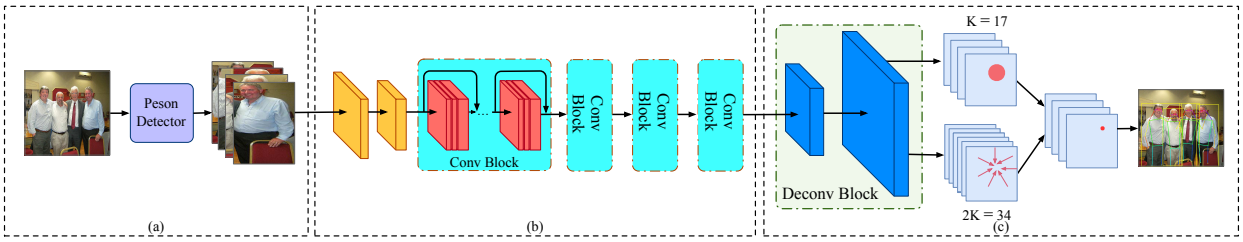

Figure 1: Offset-guided two-stages pose estimation network. It consists of three main components: (a) the person detector, (b) extracting features using ResNet, (c) the process of refinement and fusion

the box with a lower score may have higher IoU with ground truth and may be eliminated by the subsequent NMS [20] process. Therefore, a Greedy Box Generation (GBG) strategy is proposed to retain more necessary candidates. For the second stage, two branches are used to obtain the final results. The offset regression branch guides heatmap classification branch to approach the ground truth. Meanwhile, the heatmap classification branch guides offset regression branch to focus on the neighborhood of ground truth.

3.1.1 Greedy Box Generation Strategy

We adopt Mask R-CNN [12] as the person detector that achieves AP 51.7 of 80 categories detection on the COCO val2017. Different from most of the other approaches, we propose a greedy box generation (GBG) strategy as a novel alternative to Non-Maximum Suppression (NMS) [20]. It prefers to retain redundant boxes which helps us to get better pose selected by OKS+IOU NMS after pose estimation. Specifically, no filtering strategies including NMS are used in both RPN and R-CNN phase. As a result of person detection, thousands of boxes are put out as candidates. The sequential selection of candidates can be described as follows. Firstly, based on the task limitation, we filter out the boxes whose size are smaller than the minimum threshold. Then, those boxes whose confidence score is larger than 0.8 are picked out. We argue those boxes are reliable and call them equivalent ground truth (EGT). After that, the other predicted boxes who has a $IoU < 0.5$ with all EGT are eliminated. Finally, all of the rest boxes are divided into groups where every box has a $IoU >= 0.7$ with each other, and top N of each group (we use $N = 4$) are preserved. By adopting GBG strategy, we tend to keep the boxes with score relatively small but localization more accurate.

3.1.2 Offset-guided Network

In this work, we utilize ResNet [13] as the backbone of the offset-guided network. Our offset-guided architecture addresses two main problems. Firstly, in order to preserve more local details, *deconv* layers are appended for higher resolution. In our practice, two *deconv* layers are used to enlarge the feature maps by four times. Secondly, following [22], we adopt an approach combining classification and regression branches to obtain the final pose results which helps to reduce quantization errors. We denote the number of keypoints by K. A convolution layer of $K = 17$ channels is adopted to output coarse location, and a convolution layer of $2K$ channels to regress the offset for a fine position. For each predicted position x_i and each GT key point g_k, the target label for the classification head is:

$$H_c = \begin{cases} 1 & ||x_i - g_k|| <= R \\ 0 & ||x_i - g_k|| > R \end{cases} \quad (1)$$

The target label for the x-axis of offset is:

$$H_r = \begin{cases} (g_k - x_i)/R & ||x_i - g_k|| <= R \\ 0 & ||x_i - g_k|| > R \end{cases} \quad (2)$$

And the same is y-axis. The classification head considers the whole heatmaps, while the offset loss is only computed within a disk of radius R from each keypoint. Our insight is that these two heads can revise each other. The regression head helps to revise the coarse location of keypoints. The classification head helps to exclude the invalid regions, so the regression head can focus on learning offset within a small range. Besides, this heatmap-offset aggregation method outputs result in continuous space which eliminates the quantization errors. As shown in Figure 1, the OGN can be split into three stages. In experiments, the OGN dramatically improves the performance in a large range of output resolutions, especially for low resolution.

3.1.3 Inference

Inspired by [9], to make the pose estimator adapted to the boxes generated by our person detector, we mix up the predicted boxes and ground truth boxes. With this strategy, our pose estimator can adapt to the variance of box location distribution and perform better while testing. Once those ROIs are provided, the cropped areas from the original image will be sent to a single pose estimator. In our practice, ResNet is used to extract features and some *deconv* layers [29] is added to pursue higher resolution. Smooth L_1 is

used as the loss function for both classification and regression. In addition, we employ a Gaussian filter to make the output heatmaps smoother. The final results are obtained by merging two branches using an intuitive but effective fusion method.

For classification branch, each key point is predicted by a heatmap $L_k \in R_{W*H}$ (W, H is the width and height of the final heatmap respectively). The other branch is used to generate $2K$ heatmaps. Every pair of them stands for the x, y offset for the corresponding position in L_k , and these heatmaps are denoted by O_{kx}, O_{ky}. Each O_{kx} and O_{ky} have the same size with L_k. Firstly, we find the maximum score in each L_k and mark them as coarse localization (T_w, T_h).

$$(T_w, T_h) = argmax(L_k), k = 1, 2, ..., k \quad (3)$$

Then, the corresponding offsets O_{kx_w}, O_{ky_h} in O_{kx} and O_{ky} are obtained. Finally, the output can be denoted as:

$$(F_x, F_y) = (T_w + O_{kx_w}R, T_h + O_{ky_h}R) \quad (4)$$

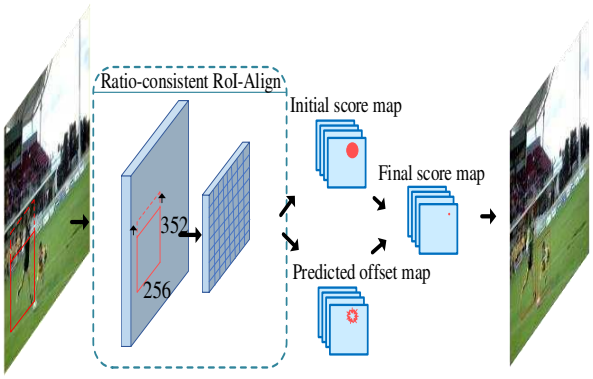

Figure 2: The framework of the offset-guided Mask R-CNN. The ROIs from RPN are firstly extended to a fixed ratio and then ROI-Align is used to extract the feature in each extended ROIs. Then a score branch and an offset branch are predicted and fused to obtain the final location of keypoints.

3.1.4 Discussion

Compared to [22], our method emphasizes simplicity and effectiveness. [22] adopts logistic loss for the classification head and Hober robust loss for the regression head, while only the Smooth L_1 loss is used for both of them in this paper. The totally different loss types in [22] introduce a hyper-parameter to keep loss balanced. In contrast, we only need to simply add the loss of the two branches together. When it comes to the process of fusion, [22] adopts Hough voting strategy while we directly select the maximum prediction. What is more, our network can still converge well without any intermediate supervision while [22] adds an extra heatmap to contribute auxiliary loss. From these perspectives, OGN can be easily transferred to other frameworks like Mask R-CNN [12]. Our approach is not only simple and intuitive, but also effective. In Section 4.1, comprehensive experiments are conducted to verify the effectiveness.

3.2. Offset-guided Mask R-CNN

Besides the above two-stages pose estimation, the effectiveness of the OGN is also evaluated on Mask R-CNN, which is an end-to-end framework producing results of detection and pose estimation simultaneously. [12] models the location of a keypoint as a one-hot mask and produces K masks for each keypoint based on the feature from ROI-Align. However, ROI-Align will output distorted feature map in different degrees if the ratios of ROIs are different, which increases the training difficulty of subsequent prediction head. And the resulting one-hot map may be less accurate due to the small resolution of the feature map.

Therefore, this paper proposes two techniques to improve the performance of human pose estimation of Mask R-CNN. The first one is to transform all human ROIs into a fixed ratio by extending the width or height of ROIs, which makes sure the ROIs fed into prediction head fall into the same distribution of ratio and improve the ability of generalization. This strategy is denoted as *ratio-consistent* in the following sections. The second one is that the human pose is predicted with a score map and an offset map. Specially, the prediction with the max response in the score map represents the coarse prediction location, and the offset map further refines it to a finer location. Here the score map is the same as the score map in single person pose estimation model mentioned in Section 3.1 and we also use Smooth L_1 loss as the optimization target. The extended Mask R-CNN framework is illustrated in Figure 2.

4. Experiments

In this section, the performance of the proposed offset-guided network is evaluated on COCO and PoseTrack dataset.

4.1. Results on COCO dataset

Experiments are firstly conducted on the COCO [18] benchmark which requires both person detection and body parts localization in uncontrolled conditions. The COCO dataset contains more than 200k images and 250k person instances splitting into train, validation and test sets. Ablation study is conducted on the validation set. To compare with other methods, we provide final results on both test-dev and test-challenge2018. The qualitative results of the COCO dataset are shown in Figure 4.

Table 1: Ablation study on the COCO *val2017* set (* indicates that flip test is used). Method (*a*) reduces the number of *deconv* layers to one in MSRA [29] published source code. Method (*b*) comes from [29]. Based on our network, method (*c*-*l*) conduct experiments on offset, GBG strategy, input size and *deconv* layers.

Method	Network	Input Size	*Deconv*	Feature Stride	Offset	GBG	AP	$AP_{0.5}$	$AP_{0.75}$	AP_m	AP_l
a	ResNet-50(MSRA)*	256x192	1	16	✗	✗	52.0	88.8	58.2	51.5	53.1
b	ResNet-50(MSRA)*	256x192	2	8	✗	✗	68.2	-	-	-	-
c	ResNet-50	256x192	2	8	✓	✗	69.7	88.2	77.2	66.0	75.8
d	ResNet-50	256x192	1	16	✓	✗	67.7	90.3	74.6	64.5	72.9
e	ResNet-50	352x256	1	16	✓	✓	70.4	90.5	76.4	67.0	75.8
f	ResNet-50	384x288	1	16	✓	✓	70.7	88.5	77.4	66.3	77.6
g	ResNet-50	512x384	1	16	✓	✓	71.7	88.7	77.7	67.0	79.0
h	ResNet-50	512x384	1	16	✓	✗	71.3	88.6	77.7	67.3	78.5
i	ResNet-50	384x288	2	8	✓	✓	71.6	89.0	78.3	67.3	78.5
j	ResNet-50	512x384	2	8	✓	✓	73.0	91.5	79.6	68.8	78.8
k	ResNet-101	512x384	2	8	✓	✓	73.8	91.6	79.6	70.0	79.7
l	ResNet-152	512x384	2	8	✓	✓	74.0	91.5	79.7	70.1	79.9

Train details Our offset-guided two-stages model is pretrained on the Imagenet [7] classification dataset. For data augmentation, random flip, rotation ($\pm 30^\circ$) and scale ($0.9 \sim 1.2$) on original image are adopted. Considering the peculiarity of multi-person pose estimation task, we use a ROI based sampling strategy to improve the model's generalization ability. Eight TITAN X GPUs and batch size of 64 are used. For every iteration, we randomly choose two images for each GPU and four ROIs for each image. The whole train process contains 22 epochs. The learning rate is 0.02 and drops twice at the 17th epoch and the 21st epoch with the decay of 0.1, SGD optimizer is used.

Test details The test is conducted on the COCO val2017, test-dev and test-challenge2018. Following our GBG strategy, all ROIs generated by detected boxes are adjusted to a fixed ratio 3:4. For post-processing, a Gaussian filter is used to smooth the heatmaps at first. Then following [5], we use the product of box score and pose score as the final score for the sorting mechanism. Finally, NMS [20] based on $IoU = 0.6$ and $OKS = 0.75$ is employed.

4.1.1 Ablation study

Ablation study is conducted on the COCO val2017 set. Offset, GBG, Resolution and network depth are considered as shown in Table 1.

1. **Offset** Network with low resolution output is of significance for resources-restricted applications due to its efficiency. From method (*a*, *d*), it can be seen that $stride = 16$ will inevitably deteriorate the performance if offset is not considered. Performance can be improved by 15.7 AP when considering offset. When $stide = 8$, our OGN method (*c*) improves MSRA baseline method (*b*) by 1.5 AP. As shown in Table 3, our offset-guided architecture also improves Mask R-CNN by 2.1 AP.

2. **GBG** From the comparison of methods (*g*, *h*), the AP can be improved by 0.4 using our GBG strategy.

3. **Resolution** Resolution is affected by input size and network stride. Comparing methods (*e*, *f*, *g*) with each other, one can find that larger network input produces better results within certain range. As input size grows, the AP increases by 0.3 and 1.0 respectively. *Deconv* layers can reduce the network stride as shown in methods (*f*, *i*) and (*g*, *j*). Similarly with [29], AP increases by 0.9 from method *f* to *i*. When adopting larger input size, our final AP can increase by 1.3 from method *g* to *j*.

4. **Network depth** Comparison of methods (*j*, *k*, *l*) exposes that performance benefits from deeper network. Changing network depth, AP can increase by 0.8 from ResNet-50 to ResNet-101 and 1.0 from ResNet-50 to ResNet-152.

4.1.2 Comparison with state-of-the-art results

The proposed OGN method participates in both COCO Keypoints 2017 and 2018 challenges. In 2017, the performance of our single model is 71.3 AP, and our final result reaches 72.8 AP on *test-dev* set when state-of-the-art is 73.0. In 2018, as shown in Table 2, our single model method, without additional training data and ensemble, achieves new state-of-the-art performance on the COCO *test-dev* set with 74.0 AP, which yields 14% relative gain compared to [22]. Comparing with the previous state-of-the-arts [29], our approach improves the results by 0.2 AP. Our result with 100k additional data and the ensemble of ResNet [13], ResNext [19], Xception [6] achieves 75.9 AP. In *test-challenge* set, our result ranks the *3rd* by 74.1 AP among COCO leaderboard when submitted.

Figure 3: Qualitative results of the PoseTrack dataset. In each frame, bounding box and pose of the human are illustrated, where the same color boxes indicate the same identity.

Table 2: Pose estimation performance with single model on the COCO *test-dev* set.

Method	AP	$AP_{0.5}$	$AP_{0.75}$	AP_m	AP_l	AR
CMU-Pose[4]	61.8	84.9	67.5	57.1	68.2	-
Mask-RCNN[12]	63.1	87.3	68.7	57.8	71.4	-
G-RMI[22]	64.9	85.5	71.3	62.3	70.0	69.7
CPN[5]	72.1	90.5	78.9	67.9	78.1	78.7
MSRA[29]	73.8	91.7	81.2	70.3	80.0	79.1
Ours-2017	71.3	91.0	78.3	67.9	76.3	74.4
Ours-2018	74.0	91.1	81.1	69.8	80.5	79.7

Table 3: Compared with Mask R-CNN on the *test-dev* set.

Method	AP	$AP_{0.5}$	$AP_{0.75}$	AP_m	AP_l	AR
Mask R-CNN[12]	63.1	87.3	68.7	57.8	71.4	-
Extended Mask R-CNN	65.2	88.2	70.9	61.8	71.7	69.8

4.2. Results on PoseTrack dataset

Experiments about pose estimation and tracking on PoseTrack dataset are also conducted. Ablation study on offset-guided Mask R-CNN is conducted. Similar tracking strategy as [11] is adopted except that the appearance information is taken into account. Specifically, we utilize the metric which integrates the spatial cue and the appearance cue. IoU is adopted to measure the spatial similarity and human Re-identification model is utilized to extract the appearance feature of the targets. Furthermore, the Euclidean distance is adopted to measure the appearance similarity.

4.2.1 Ablation study

We evaluate the proposed ratio-consistent strategy and offset-guided Mask R-CNN on PoseTrack *val2017* dataset.

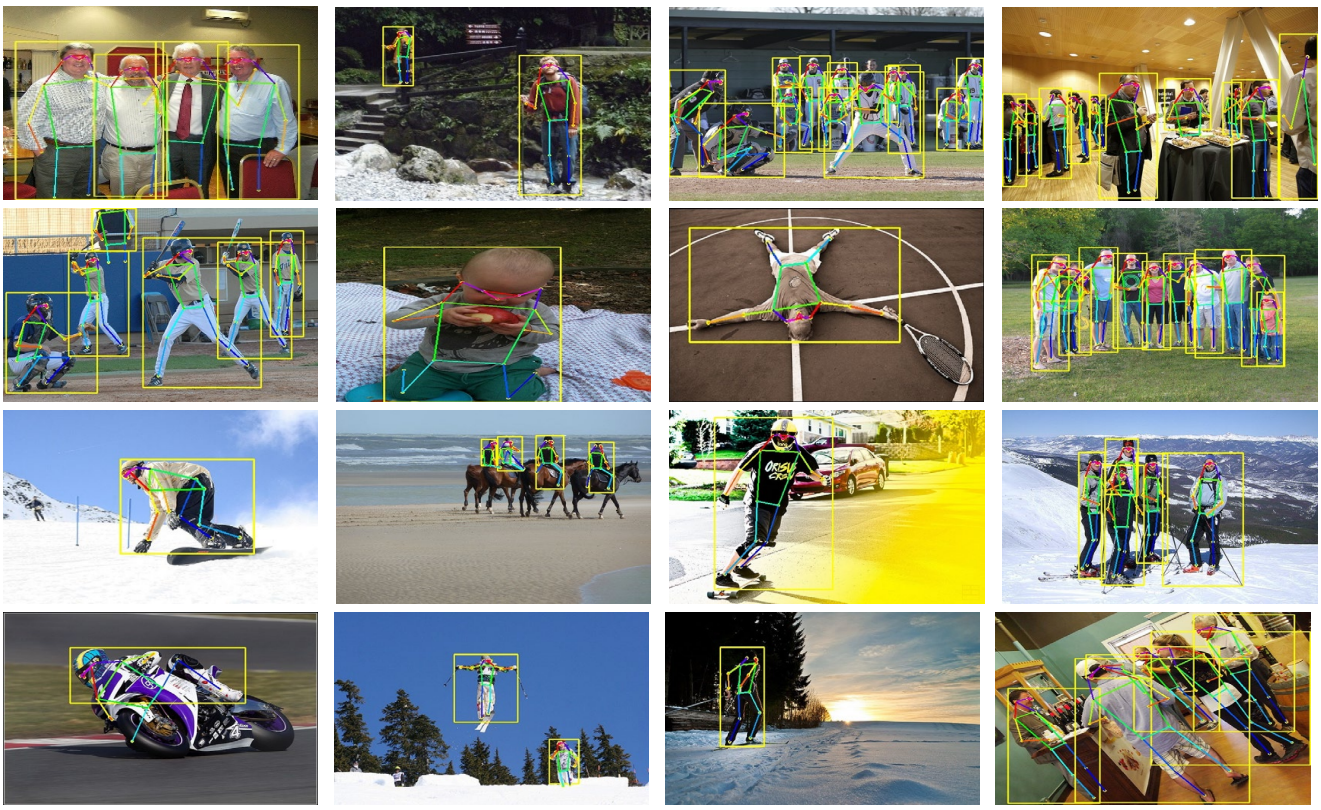

Figure 4: Qualitative results of the COCO dataset. In each frame, bounding box, keypoints and skeleton are denoted by rectangle, dot and line separately.

The results are illustrated in Table 4. We conduct experiments in three aspects including offset, ratio-consistent and the type of loss to optimize.

1. **Offset** Comparing method a with [11], the performance improvement is 3.7 mAP, which proves the effectiveness of OGN.

2. **Ratio-consistent** Similar to single person pose, the ratio of ROIs in this model is extended to 352×256. It brings another 1.6 mAP improvement comparing to method a.

3. **Loss** We regress the score maps of keypoint location and offset map with Smooth L1 loss. With this technique, 66.7 mAP is obtained. Our final results on val2017 outperform [11] by 0.8 mAP.

4.2.2 Comparison with state-of-the-art results

As shown in Table 5, without optical flow, there is an improvement of MOTA over existing best method [29] by 2.3 on PoseTrack *val2017*. If the optical flow is adopted, the MOTA improvement is 4.7. Meanwhile, Our approach obtains state-of-the-art performance on *test2017* set. The qualitative results of the PoseTrack dataset are shown in Figure 3.

Table 4: Ablation study of extended Mask R-CNN on PoseTrack *val2017* set. The backbone is ResNet-101. The result of FAIR is from [11].

Method	Loss Type	Ratio Consistent	Offset-guided Refinement	mAP Total
FAIR	softmax	✗	✗	60.6
a	softmax	✗	✓	64.3
b	softmax	✓	✓	65.9
c	Smooth L_1	✓	✓	66.7

5. Conclusion

In this paper, we revisit the heatmap-offset aggregation method for pose estimation and propose the Offet-guided network (OGN) for both two-stages approaches and Mask R-CNN. The OGN is designed to reduce errors caused by the quantization effect between network input and output. A novel alternative to NMS for two-stages network is proposed which named GBG. For offset-guided Mask R-CNN, ratio-consistent is adopted to improve the model's ability of

Table 5: Multi-person pose estimation and tracking performance on PoseTrack 2017 dataset. We adopt the same optical flow method as MSRA.

Method	Dataset	Total mAP	Total MOTA
MSRA [29]	*validation*	76.7	65.4
FAIR [11]	*validation*	64.1	55.2
PoseFlow [30]	*validation*	66.5	58.3
Ours (Mask R-CNN)	*validation*	66.7	60.7
Ours (two stages)	*validation*	75.1	67.7
Ours (two stages with optical flow)	*validation*	76.7	70.1
MSRA [29]	*test*	74.6	57.8
FAIR [11]	*test*	-	51.8
PoseFlow [30]	*test*	63.0	51.0
Ours(Mask R-CNN)	*test*	63.9	57.4
Ours(two stages)	*test*	72.6	59.2
Ours(two stages with optical flow)	*test*	74.8	61.6

generalization. State-of-the-art results are achieved on both COCO and PoseTrack dataset.

References

[1] M. Andriluka, U. Iqbal, A. Milan, E. Insafutdinov, L. Pishchulin, J. Gall, and B. Schiele. Posetrack: A benchmark for human pose estimation and tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5167–5176, 2018. 1

[2] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 1

[3] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1014–1021. IEEE, 2009. 1, 2

[4] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2, 6

[5] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun. Cascaded pyramid network for multi-person pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2, 5, 6

[6] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1800–1807, 2016. 5

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 5

[8] A. Doering, U. Iqbal, and J. Gall. Joint flow: Temporal flow fields for multi person tracking. In *British Machine Vision Conference*, 2018. 1

[9] H. Fang, S. Xie, Y.-W. Tai, and C. Lu. Rmpe: Regional multi-person pose estimation. In *IEEE International Conference on Computer Vision*, 2017. 3

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010. 1, 2

[11] R. Girdhar, G. Gkioxari, L. Torresani, M. Paluri, and D. Tran. Detect-and-track: Efficient pose estimation in videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 350–359, 2018. 1, 6, 7, 8

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *IEEE International Conference on Computer Vision*, pages 2980–2988. IEEE, 2017. 1, 2, 3, 4, 6

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 3, 5

[14] E. Insafutdinov, M. Andriluka, L. Pishchulin, S. Tang, E. Levinkov, B. Andres, and B. Schiele. Arttrack: Articulated multi-person tracking in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017. 1

[15] 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. 1, 2

[16] U. Iqbal, A. Milan, and J. Gall. Posetrack: Joint multi-person pose estimation and tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1

[17] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu. High performance visual tracking with siamese region proposal network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8971–8980, 2018. 1

[18] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 1, 2, 4

[19] J. Liu, G. Wang, L. Y. Duan, K. Abdiyeva, and A. C. Kot. Skeleton-based human action recognition with global context-aware attention lstm networks. *IEEE Transactions on Image Processing*, PP(99):1–1, 2017. 5

[20] A. Neubeck and L. Van Gool. Efficient non-maximum suppression. In *International Conference on Pattern Recognition*, volume 3, pages 850–855. IEEE, 2006. 3, 5

[21] 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. 1, 2

[22] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, volume 3, page 6, 2017. 1, 2, 3, 4, 5, 6

[23] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4929–4937, 2016. 1, 2

[24] L. Pishchulin, A. Jain, M. Andriluka, T. Thormählen, and B. Schiele. Articulated people detection and pose estimation: Reshaping the future. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3178–3185. IEEE, 2012. 1, 2

[25] M. R. Ronchi and P. Perona. Benchmarking and error diagnosis in multi-instance pose estimation. In *IEEE International Conference on Computer Vision*, pages 369–378. IEEE, 2017. 1

[26] 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. 1, 2

[27] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1653–1660, 2014. 1, 2

[28] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016. 1, 2

[29] B. Xiao, H. Wu, and Y. Wei. Simple baselines for human pose estimation and tracking. In *European Conference on Computer Vision*, 2018. 1, 2, 3, 5, 6, 7, 8

[30] Y. Xiu, J. Li, H. Wang, Y. Fang, and C. Lu. Pose flow: Efficient online pose tracking. In *British Machine Vision Conference*, 2018. 1, 8

[31] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang. Learning feature pyramids for human pose estimation. In *IEEE International Conference on Computer Vision*, volume 2, 2017. 1, 2

[32] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*, pages 487–495, 2014. 1

[33] Z. Zhu, G. Huang, W. Zou, D. Du, and C. Huang. Uct: Learning unified convolutional networks for real-time visual tracking. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1973–1982, 2017. 1

[34] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, and W. Hu. Distractor-aware siamese networks for visual object tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 101–117, 2018. 1

[35] Z. Zhu, W. Wu, W. Zou, and J. Yan. End-to-end flow correlation tracking with spatial-temporal attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1