Joint COCO and Mapillary Workshop at ICCV 2019: COCO Keypoint Detection Challenge Track

Technical Report: ByteDance HRNet

Bin Xiao, Zaizhou Gong, Yifan Lu, Linfu Wen

ByteDance AI Lab

Abstract

In this report, we present our multi-person keypoint detection system for COCO Keypoint Detection Challenge 2019. It contains three main components, which are multi-person detector, high resolution network (HRNet) for keypoint detection and pose refinement network.

As the core component, our HRNet starts from a high-resoluiton subnetwork as the first stage, gradually add high-to-low resolution subnetworks one by one to form more stages, and connect the multi-resolution subnetworks in parallel. We conduct repeated multi-scale fusions such that each of the high-to-low resolution represents reveive information from other parallel representations over and over, leading to rich high-resolution representations. As a result, the predicted keypoint heatmap is potentially more accurate and spatially more precise. With an additional pose refinement network, our final submitted result achieves an AP of 78.2 *on COCO* test-dev *set and an AP of* 75.5 *on COCO* test-challenge2019 *set respectively. The code and models have been publicly available at* `https://github.com/leoxiaobin/deep-high-resolution-net.pytorch`.

1. Overview

Figure 1 illustrates the overview of our multi-person pose estimation system for COCO Keypoint Detection Challenge 2019. Following [13, 2, 11], a two-stage top-down paradigm is applied. First, a person detector is used to localize the person in the image. Second, a core high resolution network (HRNet) [11] is used for keyoint detection. Finally, we use a pose refinement network [8] as a post processing to refine the result.

1.1. Person Detection

For person detection, by default we use a faster-RCNN [10] detector. Following [9], the backbone is a modified aligned Xception [3], equipped with deformable convolutions and deformable RoI pooling [4]. The detector achieves an AP of 61.1 for person category on COCO *test-dev* set.

1.2. High Resolution Network for Keypoint Detection

The core component in our system is the high resolution network (HRNet), which is first proposed by our recent work in [11]. Our HRNet connects high-to-low subnetworks in parallel. It mantains high-resolution representations through the whole process for spatially precise heatmap estimation. It generates reliable high-resolution representations through repeatedly fusing the representations produced by the high-to-low subnetworks. More details about HRNet are described in our recent work [11, 12].

For COCO Keypoint Detection Challenge 2019, we use HRNet-W48 as our backbone, where 48 represents the widths of hte high-resolution subnetworks in last three stages. The widths of other three parallel subnetworks are $96, 192, 384$.

1.3. Pose Refinement Network

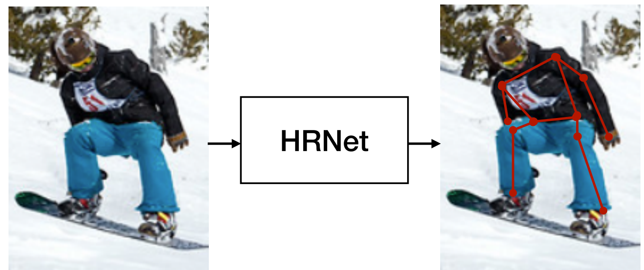
Inspired by [8], we use a refinement network as our post processing for our final submission. We use the pre-trained refinement network provided by [8].

2. Experiments

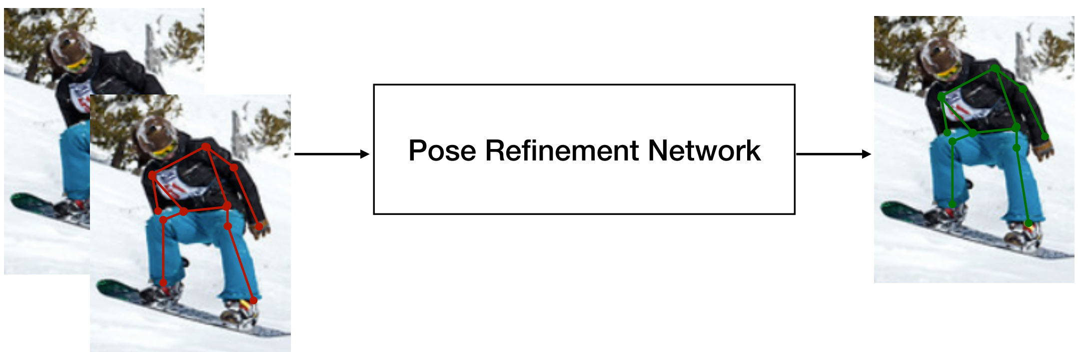
2.1. Dataset

The COCO dataset [7] contains more than $200,000$ images and $250,000$ person instances labeled with keypoints. COCO dataset [7] is split into *train/val/test-dev* sets with $57K$, $5K$ and $20K$ images respectively. An extra dataset from AI Challenger [12] is involved for training, which contains $210,000$ images and $378,374$ person intances for training. We use COCO [7] *train* set and AI Challenge *train*

(a) Person Detection



(b) HRNet for Keypoint Detection



(c) Pose Refinement

Figure 1: Overview for our multi-person pose estimation system for COCO Keypoint Detection Challenge 2019.

set to train our pose estimation models for our final submission.

2.2. Training

Our training strategy is the same as in [11]. We extend the ground truth human box in height or width to a fixed aspect ratio: $height : width = 4 : 3$. Then we crop the human box from the image, and resize to a resolution of 384×288 for training the pose estimation networks. We do data augmentation including random rotation ($[-45°, 45°]$), random scale ($[0.65, 1.35]$), random flipping and half body data augmentation [6].

Our HRNet [11] backbone network is initialized by pretraining on ImageNet classification task [5]. Adam [1] optimizer is used for training pose estimation network. The base learning rate is $1\text{e}{-3}$, and it drops to $1\text{e}{-4}$ and $1\text{e}{-5}$ at the $170th$ and $200th$ epoch respectively. There are 210 epochs in total. Mini-batch size is 32 for per GPU card. Eight GPUs are used for training.

2.3. Testing

As mentioned in Section 1, a two-stage top-down paradigm is applied. For multi-model ensemble testing, all the heatmaps generated by all the models are averaged for joint prediction. In our final submission, we used six models for model ensemble. Following the common practice in [11, 13], a quarter offset in the direction from highest response to the second highest response is used to obtain the final location. After getting the ensemble result of pose estimation, we feed the result with the input image to a pose refinement network [8] to get the final result.

2.4. Ablation Study

Table 1 shows an ablation study including using extra data for training, model ensemble and using a pose refinement network as post processing on COCO [7] *val2017* set. By default, we use the same person detector as in [13, 13]. Our baseline method (a) obtains an AP of 76.3, which is trained on COCO [7] *train2017* set with an input size of

	w/ extra data	w/ model ensemble	w/ pose refinement	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
(a)				76.3	90.8	82.9	72.3	83.4	81.2
(b)	✓			77.5	90.9	83.9	73.7	84.5	81.2
(c)	✓	✓		78.5	91.1	84.4	74.9	85.5	83.1
(d)	✓	✓	✓	78.9	91.2	84.6	75.4	85.8	83.4

Table 1: Ablation study on COCO val2017 set.

Method	Backbone	Input Size	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR	AR^{50}	AR^{75}	AR^{M}	AR^{L}
CPN* [2]	Res-Inception	384×288	73.0	91.7	80.9	69.5	78.1	79.0	95.1	85.9	74.8	84.6
Simple Base*+ [13]	Res-152	384×288	76.5	92.4	84.0	73.0	82.7	81.5	95.8	88.2	77.4	87.2
MSPN*+ [6]	4×Res-50	384×288	78.1	94.1	85.9	74.5	83.3	83.1	96.7	89.8	79.3	88.2
Our: HRNet*+	HRNet-W48	384×288	77.9	93.1	85.3	74.3	83.9	82.6	96.0	89.0	78.6	88.1
Our: HRNet*+ + refine	HRNet-W48	384×288	78.2	92.8	85.5	74.8	84.1	82.8	95.9	89.1	78.9	88.2

Table 2: Comparisons of results on COCO *test-dev2017* dataset. ”*” indicates using an ensemble model and ”+” means using external data.

384×288, using HRNet-W48 [11] as backbone. With an additioanl AI Challenger data set [12] involved for training, our method (b) achieves an AP of 77.5, which is 1.2 AP better than the baseline. And an ensemble model (c) obtains an AP of 78.5. Finally, with a pose refinement as post processing, our method (d) achieves an AP of 78.9, which has an improvement of AP by 2.5.

2.5. Results

Table 2 shows comparisons of results on COCO *test-dev2017* set. With extra data involved in training, an ensembled models of our HRNet [11] obtains an AP of 77.9. Our final submission for COCO 2019 Keypoint Detection Challenge is further refined by an pose refinement network [8], which achieves an AP of 78.2 on COCO *test-dev* set, and achieves an AP of 75.5 on COCO *test-challenge2019* set.

References

[1] Yoshua Bengio and Yann LeCun, editors. *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 2

[2] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded Pyramid Network for Multi-Person Pose Estimation. 2018. 1, 3

[3] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 1

[4] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 1

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[6] Wenbo Li, Zhicheng Wang, Binyi Yin, Qixiang Peng, Yuming Du, Tianzi Xiao, Gang Yu, Hongtao Lu, Yichen Wei, and Jian Sun. Rethinking on multi-stage networks for human pose estimation. *CoRR*, abs/1901.00148, 2019. 2, 3

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 1, 2

[8] Gyeongsik Moon, Juyong Chang, and Kyoung Mu Lee. Posefix: Model-agnostic general human pose refinement network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 3

[9] Haozhi Qi, Zheng Zhang, Bin Xiao, Han Hu, Bowen Cheng, Yichen Wei, and Jifeng Dai. Deformable convolutional networks–coco detection and segmentation challenge 2017 entry. In *ICCV COCO Challenge Workshop*, volume 15, 2017. 1

[10] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1

[11] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2, 3

[12] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, Yizhou Wang, and Yonggang Wang. AI challenger : A large-scale dataset for going deeper in image understanding. *CoRR*, abs/1711.06475, 2017. 1, 3

[13] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *The European Conference on Computer Vision (ECCV)*, September 2018. 1, 2, 3