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

Technical Report: Multi-Stage HRNet: Multiple Stage High-Resolution Network for Human Pose Estimation

Junjie Huang, Zheng Zhu, Guan Huang
Institute of Automation, Chinese Academy of Sciences, Beijing.
zhengzhu@ieee.org

Abstract

Human pose estimation are of importance for visual understanding tasks such as action recognition and human-computer interaction. In this work, we present a Multiple Stage High-Resolution Network (Multi-Stage HRNet) to tackling the problem of multi-person pose estimation in images. Specifically, we follow the top-down pipelines and high-resolution representations are maintained during single-person pose estimation. In addition, multiple stage network and cross stage feature aggregation are adopted to further refine the keypoint position. The resulting approach achieves promising results in COCO datasets. Our single-model-single-scale test configuration obtains 77.1 AP score in test-dev using publicly available training data. [1]

1. Introduction

Human pose estimation has witnessed a significant advance thanks to the development of deep learning. Motivated by practical applications in video surveillance [11], human-computer interaction [24, 30] and action recognition [26, 28, 25, 27], researchers now switch focus from single person [18, 19, 14, 22] to multi-person pose estimation in unconstrained environments [16, 7, 2, 15, 4, 21, 12, 17]. Even though research community has witnessed a significant advance, there are still challenging pose estimation problems in complex environments, such as occlusion, intense light and rare poses.

Multi-person pose estimation can be categorized into bottom-up [16, 7, 2] and top-down approaches [15, 4, 6, 21], where the latter becomes dominant participants in COCO benchmarks [13]. Bottom-up architecture based methods first detect body parts and then associate corresponding body parts with specific human instances. Top-down approaches firstly detect and crop persons from the image, then perform the single person pose estimation in the cropped person patches.

Recently, Multiple Stage Pose estimation Network (MSPN) [12] and High-Resolution Network (HRNet) [17] set new state-of-the-art performances in COCO keypoint benchmark. In this work, we design a Multiple Stage High-Resolution Network (Multi-Stage HRNet) by elegantly combining these two awesome approaches. Specifically, we follow the top-down pipelines and high-resolution representations are maintained during single-person pose estimation. To further refine the keypoint position, multiple stage network and cross stage feature aggregation are adopted.

The proposed approach achieves promising performance in COCO keypoint benchmark. Without bells and whistles, Multi-Stage HRNet achieves 79.4 and 77.1 AP score on mini-validation and test-dev, with publicly available training data and single-model-single-scale test configuration. With simple and common test augmentation and model ensemble, we obtain 80.3 and 77.7 AP score in mini-validation and test-dev, respectively.

2. Related Works

2.1. Single person pose estimation

Recently, single person pose estimation has been advanced rapidly for the development of deep convolution neural networks (CNN). DeepPose [18] firstly tries to utilize CNN in pose estimation by directly regressing the x, y coordinates of body parts. More recently, researchers choose to regress heatmaps, where each peak stands for a body part. With the continuous work of research community, novel architectures such as CPM [19] and Stacked Hourglass [14] are proposed to achieve better results.

[1] technical report. Junjie Huang and Zheng Zhu contribute equally to this work.

2.2. Multi-person pose estimation

Different from single pre-located person, multi-person pose estimation can be categorized into bottom-up [16, 7, 2] and top-down approaches [15, 4, 6, 21, 12, 17].

bottom-up Bottom-up architecture based methods adopt a different work flow, which first detect body parts and then associate corresponding body parts with specific human instances. The typical methods are DeepCut [16] and DeeperCut [7], the former adopts an integer linear programming(ILP) based method and the later improves DeepCut via utilizing image-conditioned pairwise terms. Cao et al. [2] predict heatmaps of body parts and a set of 2D vector fields of part affinities and parse them by greedy inference to generate the final results.

Top-down CPN [4] is the leading method on COCO 2017 keypoint challenge. It involves skip layer feature concatenation and an online hard keypoint mining step. [21] adopts FPN-DCN as the human detector and adds a few deconvolutional layers on single-person pose estimation network to improve the performance. Besides, Mask R-CNN [6] builds an end-to-end framework and yields an impressive performance. Recently, HRNet [17] and MSPN [12] set new state-of-the-art results on COCO keypoint detection task. HRNet maintains high-resolution representations through the whole single person pose estimation by connecting the multi-resolution subnetworks in parallel. MSPN propses a multi-stage pose estimation network with feature aggregation across different stages and coarse-to-fine supervision strategy.

3. Multi-Stage HRNet

3.1. Overall framework

The overall framework of Multi-Stage HRNet is illustrated in Figure 1. In each stage, HRNet is adopted as backbone, which is potentially more accurate and spatially more precise than high-to-low resolution networks [21, 4, 14]. Following [17], our approach maintains high-resolution representations through the whole single person pose estimation by connecting the multi-resolution subnetworks in parallel. Besides, repeated multi-scale fusions are performed, which makes each of the high-to-low resolution representations receive information from other parallel representations. Since estimating the keypoint positions by single network is difficult, multiple high-resolution networks are cascaded to refine the pose results. Feature maps yielded by each high resolution block are delivered to the corresponding positions of the next stage by performing plus operation. Different from 4 stages used in MSPN, here we adopt 2 stages network due to training efficiency and GPU memory. Following [15], we finally decode the heat map and offset map of the last stage to position the keypoints.

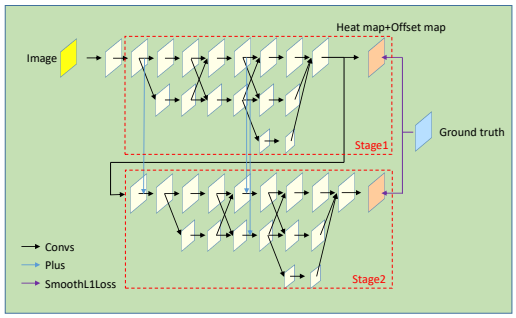

Figure 1: Overall framework of Multi-Stage HRNet.

3.2. Training details

The proposed Multi-Stage HRNet is implemented using MXNet framework. We start this work from August of 2019. There are totally 4 machines to train models, and each machine is equipped with 8 Titan RTX GPUs (24G). The multiple stage weight is initialized using single stage weight from classification task. The Adam solver with fp32 is used for training (We use fp16 to accelerate training speed at first, but it is unstable under some epochs). The initial learning rate is set to 0.001, and reduced by a factor of 10 at 110 and 140 epochs, respectively. Most model is trained with 150 epochs, and epochs of some model varies due to time limitations. Training data consists of 4 publicly available datasets, including COCO [13], MPII [1], AI Challenger [20], CrowdPose [10]. The image/instance numbers of these datasets are listed in Table 1. For data argumentation, random rotation ($[-30°, 30°]$), horizontal flipping, random resizing ($[0.8, 1.2]$), half body cropping are utilized.

Table 1: Details about training dataset, numbers are calculated by images containing pose annotations.

dataset	image number	instance number
COCO	56599	149760
MPII	17408	28800
AI Challenger	209888	377856
CrowdPose	12000	42624

3.3. Test details

For person detector, we use the HTC [3] with multi-scale test. The 80-class and person AP on mini-validation are 52.9 and 65.1, respectively. In ablation study, we adopt single-model-single-scale configuration with flipping strategy. No test argumentation or ensemble are performed. For finally ensemble, three 2-stage (i.e Multi-Stage HRNet-W48*2) are utilized and results are obtained by averaging

the position in images. For test argumentation in ensemble, rotation ($\pm 20°$) and multi-scale (3 scales) are used.

4. Experiments

In this section, we report the preliminary results on COCO mini-validation and test-dev dataset. It is noting that results may update according to further experiments.

4.1. Ablation study

In this section, the ablation study of Multi-Stage HRNet is performed and results are listed in Table 2.

Input size In mini-validation set and HRNet-w32 backbone, the AP increase from 76.3 to 77.6 when input size is from 256×192 to 384×288. For HRNet-w48 backbone, the AP steadily increases when input size gets larger. The performance in test-dev set is similar with mini-validation.

Backbones Larger backbones always bring better performance. In mini-validation set and 256×192 input size, the AP of HRNet-w32 and HRNet-w48 backbone are 76.3 and 76.9, respectively. In test-dev, the performance is 73.9 and 74.3, respectively.

Multiple stages We validate the effectiveness of multiple stages with HRNet-w48 backbone and 512×384 input size. In mini-validation and test-dev set, the improvements are 0.4 and 0.3, respectively.

Training data and ensemble More training data could boost the representation and generalization of deep learning models. In this work, we only utilize the publicly available pose data to train the Multi-Stage HRNet, which is illustrated in Table 1. As shown in Table 2, additional data could boost mini-validation and test-dev set by 1.2 and 1.3 AP, respectively. Finally, ensemble brings 0.9 and 0.6 AP for mini-validation and test-dev set.

4.2. Comparison with other methods

In this section, the performance of proposed method is compared with single-model methods on COCO test-dev set. As illustrated in Table 3, our 2-atage HRNet obtains 75.8 and 77.1 with COCO and all publicly available data respectively, which is very promising. It is worth noting that HRNet use smaller backbone and less training data.

We also compare ensemble results with methods in COCO leaderboard. As shown in Table 4, our AP is 77.7 which could rank second. The ranked first Megvii (Face++) use private data for training.

Table 4: Comparison with methods in COCO leaderboard.

Methods	Backbone	training data	AP
Ours	HRNet	Public	77.7
Megvii (Face++)	ResNet	Public+Private	78.1
MSRA	ResNet	Public	76.5
The Sea Monsters	ResNet	Public+Private	75.9
KPLab	ResNet	-	75.1
DGDBQ	ResNet	Public	74.9
ByteDance-SEU	ResNet	Public	74.2

5. Conclusion and future work

In this work, we present Multi-Stage HRNet to tackling the problem of multi-person pose estimation in images. Following the top-down pipelines, high-resolution representations are maintained during single-person pose estimation, and multiple stage network are adopted to further refine the keypoint position. Without bells and whistles, Multi-Stage HRNet achieves 79.4 and 77.1 AP score on mini-validation and test-dev, with and publicly available training data and single-model-single-scale test configuration. Due to time and GPU resources limitation, we currently adopt simple and common data augmentation during training and test. There also may be some misalignments for flip, rotation and resize. Future work may fix these misalignments and explore special augmentation strategies. Another future work may combine Multi-Stage HRNet with tracking strategy [29, 9, 31, 32, 11] for pose tracking tasks [8, 23].

References

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

[2] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 1, 2

[3] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang, et al. Hybrid task cascade for instance segmentation. In *CVPR*, pages 4974–4983, 2019. 2

[4] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun. Cascaded pyramid network for multi-person pose estimation. In *CVPR*, 2018. 1, 2, 4

[5] H. Fang, S. Xie, Y.-W. Tai, and C. Lu. Rmpe: Regional multi-person pose estimation. In *ICCV*, 2017. 4

[6] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, pages 2980–2988. IEEE, 2017. 1, 2, 4

[7] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *ECCV*, pages 34–50. Springer, 2016. 1, 2

[8] U. Iqbal, A. Milan, and J. Gall. Posetrack: Joint multi-person pose estimation and tracking. In *CVPR*, 2017. 3

[9] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu. High performance visual tracking with siamese region proposal network. In *CVPR*, pages 8971–8980, 2018. 3

[10] J. Li, C. Wang, H. Zhu, Y. Mao, H.-S. Fang, and C. Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *CVPR*, pages 10863–10872, 2019. 2

[11] P. Li, J. Zhang, Z. Zhu, Y. Li, L. Jiang, and G. Huang. State-aware re-identification feature for multi-target multi-camera tracking. In *CVPR Workshops*, 2019. 1, 3

Table 2: Ablation study of Multi-Stage HRNet in mini-validation and test-dev set.

Stages	Backbone	Input size	training data	ensemble	dataset	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
1	HRNet-w32	256×192	COCO	No	mini-val	76.3	92.6	83.7	73.8	81.9
1	HRNet-w32	384×288	COCO	No	mini-val	77.6	91.4	82.9	73.9	83.7
1	HRNet-w48	256×192	COCO	No	mini-val	76.9	90.8	82.3	73.5	83.2
1	HRNet-w48	512×384	COCO	No	mini-val	77.8	91.6	83.3	74.3	84.3
2	HRNet-w48	512×384	COCO	No	mini-val	78.2	91.8	83.7	74.7	85.1
2	HRNet-w48	512×384	all	No	mini-val	79.4	92.6	85.3	75.3	86.0
2	HRNet-w48	512×384	all	Yes	mini-val	80.3	93.1	86.4	75.9	87.2
1	HRNet-w32	256×192	COCO	No	test-dev	73.9	91.8	80.8	70.3	79.9
1	HRNet-w32	384×288	COCO	No	test-dev	75.0	92.1	83.1	71.6	81.3
1	HRNet-w48	256×192	COCO	No	test-dev	74.3	92.0	81.4	70.8	80.4
1	HRNet-w48	512×384	COCO	No	test-dev	75.5	92.2	83.8	71.8	81.6
2	HRNet-w48	512×384	COCO	No	test-dev	75.8	92.3	84.0	72.1	81.6
2	HRNet-w48	512×384	all	No	test-dev	77.1	92.5	84.2	73.8	82.9
2	HRNet-w48	512×384	all	Yes	test-dev	77.7	93.0	84.8	74.1	83.7

Table 3: Comparisons on the COCO test-dev set with most single-model configuration.

Method	Backbone	Input size	#Params	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}
Mask-RCNN [6]	ResNet-50-FPN	–	–	–	63.1	87.3	68.7	57.8	71.4
G-RMI [15]	ResNet-101	353×257	42.6M	57.0	64.9	85.5	71.3	62.3	70.0
G-RMI + extra data [15]	ResNet-101	353×257	42.6M	57.0	68.5	87.1	75.5	65.8	73.3
CPN [4]	ResNet-Inception	384×288	–	–	72.1	91.4	80.0	68.7	77.2
RMPE [5]	PyraNet	320×256	28.1M	26.7	72.3	89.2	79.1	68.0	78.6
CPN (ensemble) [4]	ResNet-Inception	384×288	–	–	73.0	91.7	80.9	69.5	78.1
SimpleBaseline [21]	ResNet-152	384×288	68.6M	35.6	73.7	91.9	81.1	70.3	80.0
HRNet-W32 [17]	HRNet-W32	384×288	28.5M	16.0	74.9	92.5	82.8	71.3	80.9
HRNet-W48 [17]	HRNet-W48	384×288	63.6M	32.9	75.5	92.5	83.3	71.9	81.5
HRNet-W48 + extra data [17]	HRNet-W48	384×288	63.6M	32.9	77.0	92.7	84.5	73.4	83.1
Ours	HRNet-W48 $*$ 2	512×384	118.2M	61.5	75.8	92.3	84.0	72.1	81.6
Ours + extra data	HRNet-W48 $*$ 2	512×384	118.2M	61.5	77.1	92.5	84.2	73.8	82.9

[12] W. Li, Z. Wang, B. Yin, Q. Peng, Y. Du, T. Xiao, G. Yu, H. Lu, Y. Wei, and J. Sun. Rethinking on multi-stage networks for human pose estimation. *arXiv preprint arXiv:1901.00148*, 2019. 1, 2

[13] 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 *ECCV*, pages 740–755. Springer, 2014. 1, 2

[14] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499. Springer, 2016. 1, 2

[15] G. Papandreou, T. Zhu, N. Kanazawa, A. Toshev, J. Tompson, C. Bregler, and K. Murphy. Towards accurate multi-person pose estimation in the wild. In *CVPR*, volume 3, page 6, 2017. 1, 2, 4

[16] 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 *CVPR*, pages 4929–4937, 2016. 1, 2

[17] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. *arXiv preprint arXiv:1902.09212*, 2019. 1, 2, 4

[18] A. Toshev and C. Szegedy. Deeppose: Human pose estimation via deep neural networks. In *CVPR*, pages 1653–1660, 2014. 1

[19] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional pose machines. In *CVPR*, pages 4724–4732, 2016. 1

[20] J. Wu, H. Zheng, B. Zhao, Y. Li, B. Yan, R. Liang, W. Wang, S. Zhou, G. Lin, Y. Fu, et al. Ai challenger: a large-scale dataset for going deeper in image understanding. *arXiv preprint arXiv:1711.06475*, 2017. 2

[21] B. Xiao, H. Wu, and Y. Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018. 1, 2, 4

[22] F. Zhang, X. Zhu, and M. Ye. Fast human pose estimation. In *CVPR*, pages 3517–3526, 2019. 1

[23] J. Zhang, Z. Zhu, W. Zou, P. Li, Y. Li, H. Su, and G. Huang. Fastpose: Towards real-time pose estimation and tracking via scale-normalized multi-task networks. *arXiv preprint arXiv:1908.05593*, 2019. 3

[24] Z. Zhang. Microsoft kinect sensor and its effect. *IEEE multimedia*, 19(2):4–10, 2012. 1

[25] J. Zhu, Z. Zhu, and W. Zou. End-to-end video-level representation learning for action recognition. In *ICPR*, pages 645–650. IEEE, 2018. 1

[26] J. Zhu, W. Zou, L. Xu, Y. Hu, Z. Zhu, M. Chang, J. Huang, G. Huang, and D. Du. Action machine: Rethinking action recognition in trimmed videos. *arXiv preprint arXiv:1812.05770*, 2018. 1

[27] J. Zhu, W. Zou, and Z. Zhu. Two-stream gated fusion convnets for action recognition. In *ICPR*, pages 597–602. IEEE, 2018. 1

[28] J. Zhu, W. Zou, Z. Zhu, and Y. Hu. Convolutional relation network for skeleton-based action recognition. *Neurocomputing*, 2019. 1

[29] Z. Zhu, G. Huang, W. Zou, D. Du, and C. Huang. Uct: Learning unified convolutional networks for real-time visual tracking. In *ICCV Workshops*, pages 1973–1982, 2017. 3

[30] Z. Zhu, H. Ma, and W. Zou. Human following for wheeled robot with monocular pan-tilt camera. *arXiv preprint arXiv:1909.06087*, 2019. 1

[31] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, and W. Hu. Distractor-aware siamese networks for visual object tracking. In *ECCV*, pages 101–117, 2018. 3

[32] Z. Zhu, W. Wu, W. Zou, and J. Yan. End-to-end flow correlation tracking with spatial-temporal attention. In *CVPR*, 2018. 3