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

Technical Report: Towards Good Practices for Multi-Person Pose Estimation

Dongdong Yu[†], Kai Su[†], Changhu Wang
ByteDance AI Lab, China

Abstract

Multi-Person Pose Estimation is an interesting yet challenging task in computer vision. In this paper, we conduct a series of refinements with the MSPN and PoseFix Networks, and empirically evaluate their impact on the final model performance through ablation studies. By taking all the refinements, we achieve 78.7 on the COCO test-dev dataset and 76.3 on the COCO test-challenge dataset.

1. Introduction

Multi-Person Pose Estimation is a fundamental yet challenging problem in computer vision. The goal is to locate body parts for all persons in an image, such as keypoints on the arms, torsos, and the face. It is important for many applications like human re-identification, human-computer interaction and activity recognition.

The tremendous development of deep convolution neural networks [2] bring huge progress for multi-person pose estimation. Existing approaches can be roughly classified into two frameworks, i.e., top-down framework [3, 7, 5, 8] and bottom-up framework [6]. The former one first detects all human bounding boxes in the image and then estimates the pose within each box independently. For example, Multi-Stage Pose estimation Network (MSPN) [3] adopts the ResNet-50 through multi stages based on repeated down and up sampling steps. PoseFix Network [5] is a human pose refinement network that refines a estimated pose from a tuple of an input image and a pose. The latter one first detects all body keypoints independently and then assembles the detected body joints to form multiple human poses. For example, Associate Embedding [6] designs the network to simultaneously estimate the keypoints detection and group heatmaps, instead of the multi-stage pipelines.

In this paper, we follow the top-down pipeline and conduct a series of refinements based on MSPN and PoseFix Networks and evaluate their impact on the final model performance through ablation studies. Finally, we achieve 78.7 on the COCO test-dev dataset and 76.3 on the COCO test-challenge dataset.

2. Method

To handle the multi-person pose estimation, we follow the top-down pipeline. First, a human detector is applied to generate all human bounding boxes in the image. Then we apply pose estimation network to estimate the corresponding human pose.

The MSPN network adopts the ResNet-50 as the backbone of the encoder and decoder. In our work, we propose a new backbone, named Refine-50, which can well handle the scale variant cases.

3. Experiments

3.1. Datasets

The training datasets include the COCO train2017 dataset [4] (includes 57K images and 150K person instances) and all the AI Challenge dataset [1] (includes 400K person instances). For the AI Challenge dataset, we only use the same annotated keypoints as the COCO train2017 dataset for the training. The final results are reported on the COCO test-dev dataset and the COCO test-challenge dataset.

3.2. Results

3.2.1 Ablation Study

In this subsection, we will step-wise decompose our model to reveal the effect of each component. In the following experiments, we evaluate all comparisons on the COCO val2017 dataset. We use 4x stage for both MSPN network and our network.

Effect of Backbone Different with MSPN, we replace the ResNet-50 with our Refine-50. As shown in Table 1, we do the experiment with official MSPN code, the AP is 74.7,

[†] Equal contribution.

Table 1: Results with different backbones on COCO val2017 dataset.

Backbone	GT Box	Detection Box
ResNet-50(Result from Paper)	76.5	75.9
ResNet-50(Our implement from github)	76.2	74.7
Refine-50(Ours)	77.7	76.0

Table 2: Results with different resolution on COCO val2017 dataset.

Resolution	GT Box	Detection Box
256x192	77.7	76.0
384x288	78.9	77.5

Table 3: Results with different training datasets on COCO val2017 dataset.

Training Dataset	GT Box	Detection Box
COCO train2017	78.9	77.5
COCO and AI	80.2	78.5

Table 4: Results of our model on COCO2017 test-dev and test-cha dataset.

BackBone	Development set	Challenge set
Refine-50	78.0	76.0
Refine-50+PoseFix	78.7	76.3

which is obvious lower than the MSPN's result from their paper. After replacing the ResNet-50 with our Refine-50, the AP is improved from 74.7 to 76.0. Detection box is provided from the MSPN paper.

Effect of Image Resolution By using a larger resolution of input image, the AP performance is improved from 76.0 to 77.5, as shown in Table 2.

Effect of Extra Dataset Besides, we also use extra dataset(AI Challenge Dataset) for training. As shown in Table 3, the AP is improved from 77.5 to 78.5 by using the extra dataset.

3.2.2 Development and Challenge Results

In this subsection, we ensemble three Refine-50 models for the pose estimation. As shown in Table 4, the AP of test-dev is 78.0, and the AP of test-cha is 76.0. After using PoseFix, the AP of test-dev can be improved to 78.7 and the AP of test-cha can be improved to 76.3.

4. Conclusion

In this work, we conduct a series of refinements with the MSPN and PoseFix Networks, and empirically evaluate their impact on the final model performance through ablation studies. By taking all the refinements, we achieve 78.7 on the COCO test-dev dataset and 76.3 on the COCO test-challenge dataset.

References

[1] AI-Challenge. Ai challenge keypoints. https://challenger.ai/competition/keypoint/subject. 1

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[3] 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. *arXiv preprint arXiv:1901.00148*, 2019. 1

[4] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[5] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Posefix: Model-agnostic general human pose refinement network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7773–7781, 2019. 1

[6] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *Advances in Neural Information Processing Systems*, pages 2274–2284, 2017. 1

[7] Kai Su, Dongdong Yu, Zhenqi Xu, Xin Geng, and Changhu Wang. Multi-person pose estimation with enhanced channel-wise and spatial information. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5674–5682, 2019. 1

[8] Dongdong Yu, Kai Su, Jia Sun, and Changhu Wang. Multi-person pose estimation for pose tracking with enhanced cascaded pyramid network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 1