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

Technical Report: HintPose

Sanghoon Hong* Hunchul Park
Kakao Brain
Seongnam, Gyeonggi, South Korea
{sanghoon.hong, robert.p}@kakaobrain.com

Jonghyuk Park Sukhyun Cho Heewoong Park
Seoul National University
Seoul, South Korea
{chico2121, chosh90, hee188}@snu.ac.kr

Abstract

Most of the top-down pose estimation models assume that there exists only one person in a bounding box. However, the assumption is not always correct. In this technical report, we introduce two ideas, instance cue and recurrent refinement, to an existing pose estimator so that the model is able to handle detection boxes with multiple persons properly. When we evaluated our model on the COCO17 keypoints dataset, it showed non-negligible improvement compared to its baseline model. Our model achieved 76.2 mAP as a single model and 77.3 mAP as an ensemble on the test-dev set without additional training data. After additional post-processing with a separate refinement network, our final predictions achieved 77.8 mAP on the COCO test-dev set.

1. Introduction

Most of the top-down pose estimation models, such as HRNet[10], CPN[3], Mask R-CNN[4] generate predictions assuming there exists only one person in one person detection box. However, multiple persons can reside in a box in some cases, as shown in Fig 1, and it is difficult to say which one is a dominant target person. It becomes more difficult as a model processes cropped images from expanded detection boxes to capture context information. This phenomenon makes single-person pose estimation as an ill-posed problem, in which there exist multiple solutions.

To alleviate this issue, we introduce two ideas; the first idea is to add *instance cue* on input which specifies a target person in a box, and the other is to design a recurrent network so that a model can refine its predictions using the outputs from previous hops as a hint for a target person.

Figure 1: An example of two overlapping persons in one bounding box.

2. HintPose

Figure 2 shows our network structure. We adopt HRNet[10] as a baseline architecture since, to the best of our knowledge, it is the state-of-the-art model with open-sourced codes. Then we add an input refinement block so that the model can handle an external instance cue for a target person. We also create a feedback connection from the output of the network to the intermediate feature map so that the model can refine its outputs using its previous predictions. For each of both modifications, we just add two simple convolutional layers with a residual connection to reuse ImageNet pre-trained models provided by the official HRNet respository[1].

*Corresponding author

[1]https://github.com/leoxiaobin/deep-high-resolution-net.pytorch

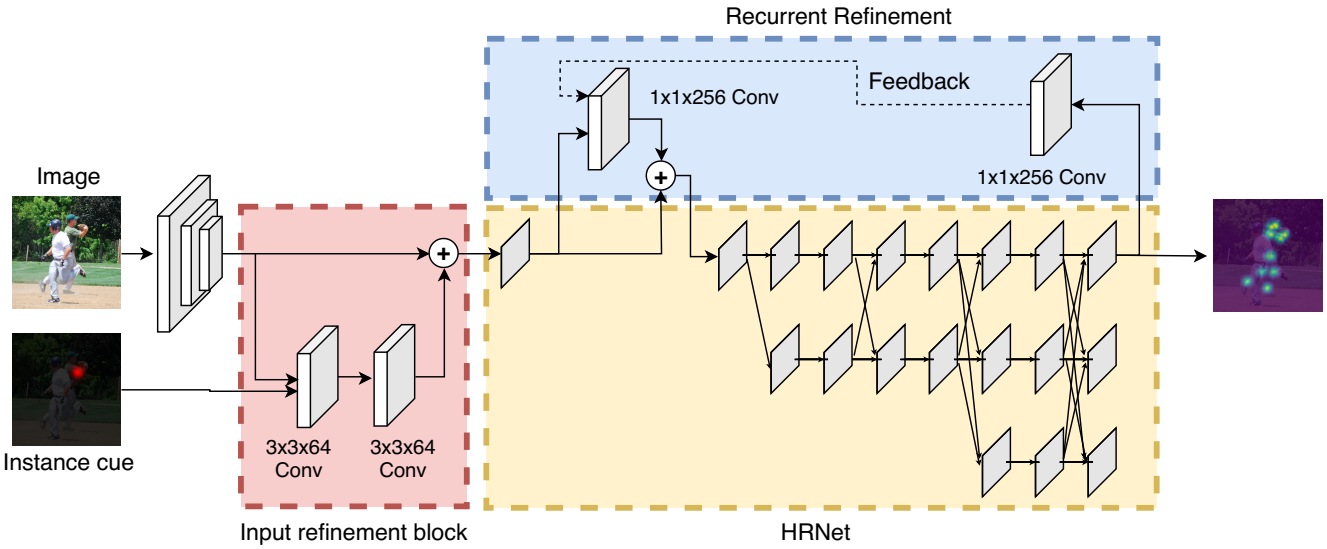

Figure 2: HintPose Network Architecture

2.1. Instance Cue

As mentioned earlier, there may exist two or more persons in a bounding box from annotated data or a detection box from an object detector. To specify a target person explicitly, a cropped image and an instance cue embedding are fed into our network. The embedding is a single channel Gaussian heatmap which has a peak located on a target person.

The input refinement block in HintPose aggregates image features and instance cue embedding and updates feature maps with element-wise summation.

Instance cues can be derived from ground-truth keypoints or instance segmentation maps during training time. At inference time, they can be generated from the outputs of other instance segmentation or keypoints estimation models, or it is also possible to train another simple network to predict them directly.

2.2. Recurrent Refinement

In addition to providing an external instance cue, it is also possible to use the outputs of the model itself as a hint for a target person. We adopted the structure of Feedback Network[7] and designed our model to have a recurrent connections so that it can refine its outputs using its previous predictions.

We added two 1×1 convolutional blocks onto the baseline network; one is to update feature maps using information from the previous output and another is to extract meaningful information for the next hop. We built the connection to feed back onto features after `layer1` of HRNet so that the improved features can be processed in all the different scales of HRNet and have smaller memory footprint.

3. Experiments

3.1. Training & Evaluation Details

Training While training the network, we generated instance cues by randomly selecting a joint among ground-truth joints and augmenting its x, y position. For a model with recurrent refinement, the model is evaluated for three hops, and all of its prediction outputs are used to compare with a ground-truth heatmap and to compute mean squared error.

We trained our models with the COCO training set only and used the same hyper-parameters provided by the official HRNet repository.

Evaluation We used MMDetection toolbox[2] and Hybrid Task Cascade (HTC)[1] + HRNetV2p-W48 model to generate detection boxes on the COCO17 validation and test sets. Its detection accuracy is 47.0 mAP (60.5 mAP for 'person' category) on the COCO17 validation set. We ignored bounding boxes smaller than 32×32.

To generate instance cues during an evaluation, we generated image-level joint heatmaps using MultiPoseNet[5] and considered local peaks in each detected bounding box as instance cues. When there are multiple cues in a bounding box, the same cropped image are fed into the model multiple times with each cue.

For models with recurrent refinement, the output heatmaps after three hops are used to compute the final predictions.

Other hyper-parameters and post-processing, including person scoring and OKS-base NMS, are kept the same with the original HRNet.

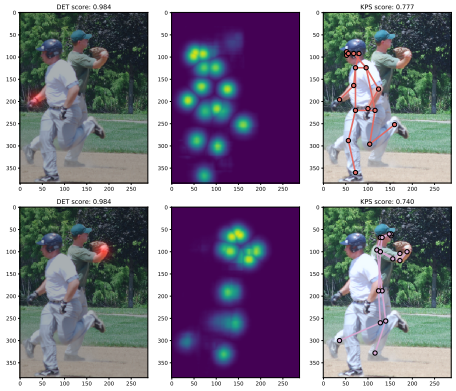

Figure 3: Two different predictions results in one person detection box. Images on the first column show cropped image and instance cue embedding. Images on the second and last column describes predicted heatmaps and skeletons.

Method	AP	AR
HRNet (Baseline)	76.7	81.2
Instance cue	77.1	81.5
Recurrent refinement	78.0	82.3
I.C. + R.R.	**78.1**	**82.6**

Table 1: COCO Validation results. 'I.C. + R.R.' stands for a model with both instance cue and recurrent refinement.

Method	AP	AR
HRNet (Baseline)	75.4	80.2
Instance cue	75.4	80.2
Recurrent refinement	76.2	81.0
I.C. + R.R.	76.2	81.0
Ensemble model	77.3	81.8
Ensemble model + PoseFix	**77.8**	**82.2**

Table 2: COCO Test-dev results. 'I.C. + R.R.' stands for a model with both instance cue and recurrent refinement.

3.2. COCO Keypoints Detection

We evaluated our models with the COCO[8] 2017 Keypoints validation and test-dev sets.

Validation set Figure 3 shows two different predictions from the same input image with two different instance cues. The predictions of our model vary as instance cue moves from one person to another.

When they are evaluated on the COCO validation set, Both *Instance Cue* and *Recurrent Refinement* improved the performance of our pose estimation model. Moreover, the improvement increased as the modifications are applied together.

Test-dev set Our model showed a significant improvement of +0.8 mAP compared to the baseline HRNet. Moreover, when we ensemble 6 different models[2], our models achieved 77.3 mAP. We also refined our predictions with PoseFix[9] and the final predictions achieved 77.8 mAP on the COCO test-dev set.

One observation is that the improvements from Instance cue are not as meaningful as those on the COCO validation set. We hypothesized that the number of crowded bounding boxes is less significant on the test-dev set.

4. Discussions & Future Works

In this technical report, we introduced two different methods to handle multiple, overlapped person instances in one bounding box. Our modifications can be applied to existing top-down pose estimation models by adding a couple of convolutional blocks. When we evaluated our model on the COCO keypoints dataset, we observed non-negligible performance improvements thanks to our HintPose method.

Our future work will include evaluating our model on other datasets and figuring out how our method performs when it is applied to other top-down models. It is known that crowded scenes are not dominant in the COCO Keypoint dataset[6]. Therefore, we expect that more significant improvements can be observed when our model is evaluated on a dataset with more occlusions such as CrowdPose[6].

It is also possible to improve our model with a better learning strategy as [7] showed that curriculum learning is vital to train its feedback network structure. Another way to improve our model can be to use a different type of instance cue, such as segmentation maps.

References

[1] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. Hybrid task cascade for instance segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[2] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 2

[3] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *The IEEE Conference*

[2] Averaged heatmaps of the followings are used for post-processing; two models with instance cue and recurrent refinement, two models with instance cue only and two models with recurrent refinement only.

on Computer Vision and Pattern Recognition (CVPR), 2018. 1

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[5] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. MultiPoseNet: Fast multi-person pose estimation using pose residual network. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[6] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. *arXiv preprint arXiv:1812.00324*, 2018. 3

[7] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. Feedback network for image super-resolution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2, 3

[8] 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 (ECCV)*, 2014. 3

[9] 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. 3

[10] 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)*, 2019. 1