# –Supplementary Material– Whole-Body Human Pose Estimation in the Wild

Sheng Jin[1,2][0000−0001−5736−7434], Lumin Xu[3,2], Jin Xu[2], Can Wang[2], Wentao Liu[2][0000−0001−6587−9878], Chen Qian[2], Wanli Ouyang[4], and Ping Luo[1]

[1] The University of Hong Kong [2] SenseTime Research
[3] The Chinese University of Hong Kong [4] The University of Sydney
{jinsheng, xulumin, wangcan, liuwentao, qianchen}@sensetime.com
wanli.ouyang@sydney.edu.au, pluo@cs.hku.hk

## 1 Annotation Details

The annotation of face/hand keypoints in our COCO-WholeBody dataset follows semi-automatic methodology. Firstly, face/hand bounding boxes are annotated manually. Secondly, we utilize a face model and a hand model, which are trained on large-scale face datasets and hand datasets respectively, to pre-annotate the face and hand keypoints. Next, manual correction of the face/hand keypoints is conducted. Foot keypoints are directly manually labeled. Note that, quality inspections are conducted in every step.

**Face and hand bounding box:** To ensure the quality of face/hand bounding boxes, well-defined standards are followed. **Face bounding box** is labeled only if the box is bigger than 8 pixels and the rotation angle of the face is less than 100° from the frontal view. As for some special cases, faces of real persons in photos, posters, and clothes are labeled but faces of sculptures, models, cartoons, paintings, and animals are not. The face bounding box is defined as the minimal bounding rectangle of the face keypoints. Quality inspections are conducted by another group of annotators and bounding boxes whose positions are inaccurate are re-annotated. **Hand bounding box** is labeled when the hand image is vivid and the position of the hand keypoints can be well-determined. The box is regarded as invalid if the corresponding hand is severely occluded or part of the hand is out of the image. Special case settings follow those of face bounding box and independent quality inspections are conducted. Examples of face/hand bounding boxes are shown in Fig. 1, where only the green boxes meet our annotation requirements. More visualization results for bounding boxes are demonstrated in Fig. 2 Line#1. We have three types of bounding boxes, *i.e.* body (green), face (purple), left hand (blue) and right hand (red).

**Face Keypoints:** We apply the 68-joint face model [9] as shown in Fig. 1(b). A few occluded keypoints may be estimated by annotators if most keypoints are visible in the image. In Fig. 2, Line#2 and Line#3 visualize more examples of the face keypoint annotations.

**Hand Keypoints:** Self-occlusion is very common for hand keypoints. As a result, the annotation for hand keypoints requires trained experts and enormous

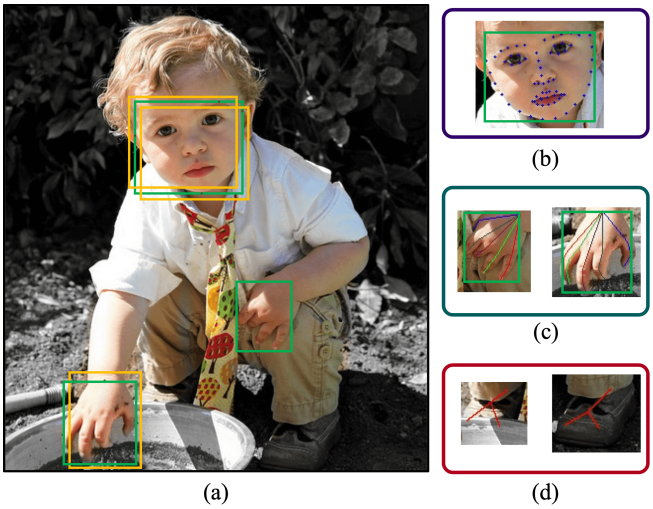


**Fig. 1.** Face/hand bounding box annotation. Bounding boxes should tightly enclose all the keypoints. Positive (green) and negative (orange) cases are shown.

workload although pseudo labels are given. We use 21-joint hand model [10] and annotate quite a lot of challenging cases. Annotation is shown in Fig. 1(c) and more examples are visualized in Fig. 2, where Line#4 and Line#5 visualize some examples of the hand keypoint annotations for various hand poses.

**Foot Keypoints:** Six foot keypoints are defined following [1]. The order in the annotation file is as follows: left big toe, left small toe, left heel, right big toe, right small toe, and right heel. The keypoints are defined in the inner center rather than on the surface to fit in images in different views. Qualitative examples are shown in Fig. 1(d).

## 2 Baseline Implementation Details

We used the official codes to reproduce existing methods. We keep all training parameters (e.g. input size, #iterations, learning rate, and so on) the same, except #keypoints (# means the number of). We also trained all the existing methods on the original 17-keypoint COCO dataset and verified that our re-implementation is the same as the original papers. For fair comparisons, all experimental results are obtained with single-scale testing. The implementation details of the baseline methods we used in the experiments are listed as following:

**OpenPose Whole-body System** [1] is a Multi-Network whole-body pose estimation system, which consists of a body keypoint model, a facial landmark detector and a hand pose estimator. We reimplement the approach by train-

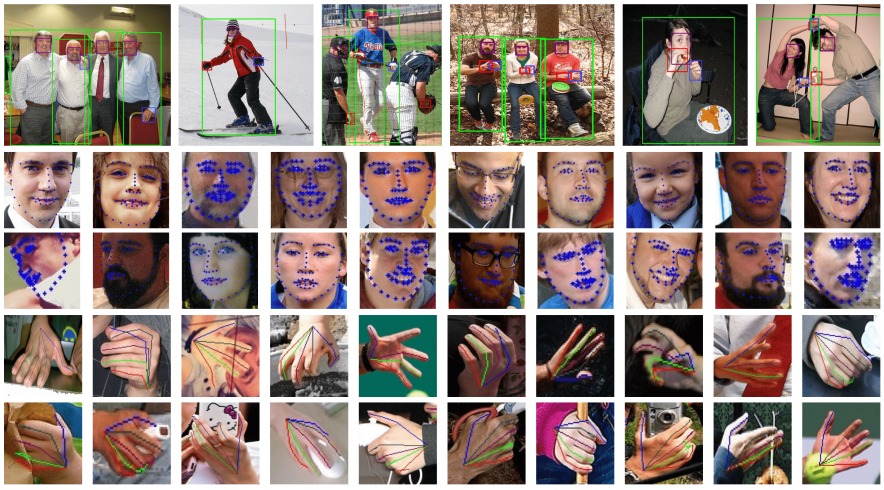

**Fig. 2.** Annotation examples. Line #1: We use different colors to distinguish different types of bounding boxes, *i.e.* body (green), face (purple), left hand (blue) and right hand (red). Line #2 and Line#3: Face keypoints. Line #4 and Line#5: Hand keypoints.

ing these models on COCO-WholeBody dataset separately based on the official training codes [1].

**Single-Network Whole-body Pose Estimation** [3] is a recently proposed method for whole-body pose estimation. We follow [3] and retrain the whole-body keypoint estimator [2] in our COCO-WholeBody dataset. The number of keypoints is 133, and the number of PAFs is 134 as we designed a tree structure except for the two loops around the lips. Face, hand and foot keypoints are connected to the corresponding nearest body keypoints. Following [3], we applied 3 stages for PAF and 1 stage for confidence maps. We use a batch size of 10 images in each GPU and Adam optimization with an initial learning rate of 1e-3 to train the model.

**Part-affinity Fields (PAF)** [2] is also re-implemented for the whole-body pose estimation task based on the open-source codes [3]. The settings of PAFs and confidence maps are the same as Single-Network [3] and CPM [12] network is used as its backbone. We use SGD with an initial learning rate of 1 to train the model. Note that, the direction of limb (or value of the affinity fields) is calculated in the image scale before down-sampling, see Fig. 3. Therefore, for most tiny hands and faces, the PAF prediction and keypoint grouping will not be affected.

**Associative Embedding (AE)** [6] learns to group keypoints by associative embedding, which is flexible in terms of various numbers of keypoints to predict.

[1] https://github.com/CMU-Perceptual-Computing-Lab/openpose
[2] https://github.com/CMU-Perceptual-Computing-Lab/openpose_train
[3] https://github.com/tensorboy/pytorch_Realtime_Multi-Person_Pose_Estimation

**Fig. 3.** Visualizations of Part-affinity Fields.

The official open-source codes [4] are used in our implementation. We use the 4-stacked hourglass backbone and follow the same training settings as in [6] in our experiments.

**HRNet** [11] is the recent state-of-the-art model for the task of multi-person human pose estimation. We retrain the model [5] to fit for the whole-body pose estimation task by directly adding the number of keypoints to 133. For fair comparisons, we choose HRNet-w32 as the backbone in the experiments. Note that this model can be viewed as the single-stage alternative of our multi-stage ZoomNet. The comparison between HRNet and ZoomNet demonstrates the effectiveness of the multi-stage keypoint localization.

## 3 ZoomNet Implementation Details

We use 2D gaussian confidence heatmaps with $\sigma = 3$ to encode the keypoint locations. The sum of squared error (SSE) loss function between the predicted heatmaps and the ground truth heatmaps is used for training both corner keypoints and body keypoints. The losses of different body parts (body, face, hand, and feet) are summed up with the same loss weight.

We follow the same setting as HRNet [11] to use data augmentation with random scaling ([-35%, 35%]), random rotation ($[-45°, 45°]$) and flipping. BodyNet and FaceHead/HandHead are first pre-trained separately and then end-to-end finetuned as a whole for 120 epochs in total. ZoomNet is trained on 8 GPUs with a batch size of 32 in each GPU. We use Adam [4] with the base learning rate of 1e-3, and decay it to 1e-4 and 1e-5 at the 80th and 100th epochs respectively.

## 4 Analysis

**Experiments on Foot Keypoint Dataset** Cao *et al.* released the first human foot dataset [1] (COCO-foot), which extends COCO [5] dataset with 15k foot annotations. We also evaluate our proposed ZoomNet on COCO-foot dataset

[4] https://github.com/princeton-vl/pose-ae-train
[5] https://github.com/leoxiaobin/deep-high-resolution-net.pytorch

**Table 1.** Body-foot AP on COCO-foot benchmark [1]. Some results are copied from [3]. Our proposed ZoomNet outperforms SN significantly.

| Method | Body AP | Foot AP |
|---|---|---|
| Body-foot OpenPose (multi-scale) [1] | 65.3 | 77.9 |
| Body-foot SN (multi-scale) [3] | 66.4 | 76.8 |
| Body-foot ZoomNet | **75.4** | **84.7** |

**Table 2.** Effectiveness of joint learning.

| Method | Body AP | Foot AP | Face AP | Hand AP | WholeBody AP |
|---|---|---|---|---|---|
| joint training | 0.743 | 0.798 | 0.623 | 0.401 | 0.541 |
| reusing features | 0.745 | 0.796 | 0.609 | 0.393 | 0.539 |
| fully independent | 0.745 | 0.796 | 0.623 | 0.419 | 0.543 |

and directly compare with OpenPose [1] and SN [3] in Table 1. We find that our proposed ZoomNet outperforms SN significantly.

### 4.1 Experiments about joint learning.

In Table 2, we explore the effectiveness of joint training of BodyNet, FaceHead and HandHead in ZoomNet. We compare (1) joint training, (2) reusing features, and (3) fully independent face/hand detectors. Joint learning improves over "reusing features" on the performance of face (0.623 vs 0.609) and hand (0.401 vs 0.393) for more efficient feature learning. Fully independent method requires two additional models with increased complexity, but achieves limited gain (0.543 vs 0.541).

**Face/Hand Bounding Box Detection** In this section, we compare the results of face and hand bounding box detection. Compared to human body detection, detecting small objects such as face and hands are more challenging, since they only occupy a relatively small area in the whole image. General detection approaches such as Faster RCNN [7] usually treat body/face/hands as normal objects and detect all of them at once. However, note that the human body is inherently a multi-level structure, where the face/hands are low-level objects of the high-level human body. Intuitively, the location of the human body will guide the detection of face/hands. Common detection methods usually ignore the inherent correlation between the human body and the face/hands, which will lead to inferior performance. To deal with the scale variance problem, ZoomNet first locates all the person bounding boxes from the image and then detects the face and hands in each bounding box. This multi-level design enables the model to focus on the potential location of the sub-objects and ignore the disturbing background. Therefore, it is beneficial for detecting small sub-objects such as face and hands. As shown in Table 3, ZoomNet outperforms the Faster RCNN

**Table 3.** Face/hand bounding box detection results on our COCO-WholeBody benchmark. Our proposed ZoomNet outperforms Faster RCNN [7] because of its multi-level design which better handles the scale variance.

| Method | face | | lefthand | | righthand | |
|---|---|---|---|---|---|---|
| | AP | AR | AP | AR | AP | AR |
| Faster RCNN [7] | 0.439 | 0.712 | 0.266 | 0.440 | 0.262 | 0.430 |
| ZoomNet | **0.582** | **0.728** | **0.349** | **0.463** | **0.356** | **0.458** |

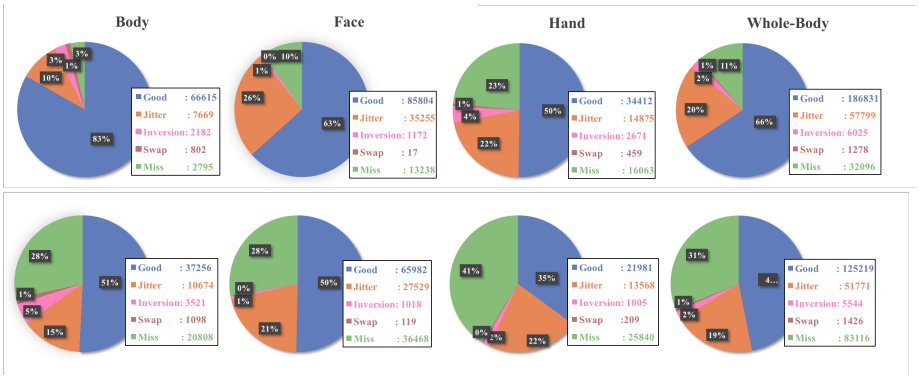


**Fig. 4.** Localization error comparison between our proposed ZoomNet (top) and Single-Network [3] (bottom). ZoomNet significantly outperforms Single-Network in the distribution of the localization error for body, face, hand and whole-body.

model by a large margin, demonstrating the effectiveness of our multi-level object detection.

**Error Analysis** In this section, we provide a more detailed error analysis for ZoomNet and Single-Network [3]. The breakdown of errors over different body parts is shown in Fig. 4. We follow [8] to define four types of localization errors, *i.e.* Jitter, Miss, Inversion, and Swap. *Jitter* means small error around the correct keypoint location, while *Miss* means the detection is not within the proximity of any ground truth body part. *Inversion* means the joint type of detected keypoint is wrong. *Swap* means the detected keypoint is grouped to a wrong person instance. On the other hand, *Good* indicates correct prediction.

We use the pie chart to show the distribution of the localization errors for the body, face, hand, and whole-body. *Miss* is the major error for all parts, and the accuracy of the hand keypoints is lower than that of the body and face keypoints. Also, ZoomNet has a higher proportion of *Good* keypoints than Single-Network.

**Size Sensitivity** In this section, we analyze the sensitivity of our proposed ZoomNet to different person sizes. To this end, we separate the COCO-WholeBody dataset into four size groups: *i.e.* medium (M), large (L), extra-large (XL) and

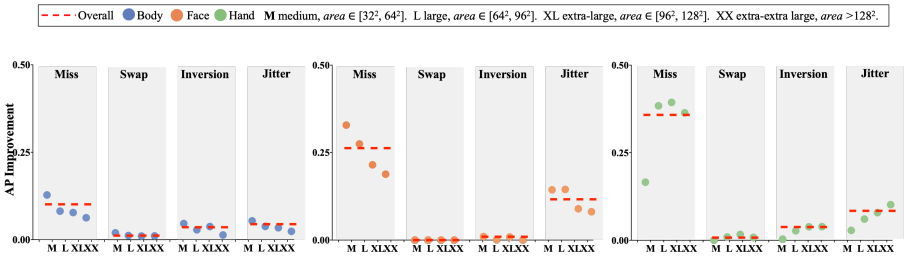


**Fig. 5.** The AP improvement obtained by correcting each type of error (including Miss, Swap, Inversion, and Jitter) for body, face, and hand separately. We use the dashed red lines to indicate performance improvement over all the instance sizes.

extra-extra large (XX). We follow [8] to use the area of the person to measure the person size, M for $area \in [32^2, 64^2]$, L for $area \in [64^2, 96^2]$, XL for $area \in [96^2, 128^2]$, and XX for $area > 128^2$. In Fig. 5, we show the AP improvement obtained after correcting each type of localization error. We find that for body and face keypoint localization, the performance can be significantly improved by correcting small-scale human poses, especially the Missing error. For hand pose estimation, errors impact performance more on larger instances. For larger-scale instance, instead of only estimating the rough position, more accurate keypoint localization is required. However, due to the frequent motion blur and severe occlusion (interaction with objects), it is still very challenging to estimating the hand poses of large instances.

**Qualitative Analysis** Fig. 6 shows the qualitative evaluation results of our approach, and Fig. 7 qualitatively compares the results of ZoomNet, OpenPose [1] and Single-Network [3]. Both of them show the capacity of our proposed ZoomNet in handling challenges including occlusion, close proximity, and small scale persons. We find that our ZoomNet significantly outperforms the previous state-of-the-art method [3], especially for face/hand keypoints. First, we observe that compared to these bottom-up approaches, ZoomNet better handles the small scale problem of human instances (see Line#1,2,3). Second, we find that the grouping of OpenPose [1] and Single-Network [3] is sometimes erroneous due to lack of human body constraints (see Line#4). Third, ZoomNet is generally better at localizing the hand/face keypoints with occlusion, pose variations, and small scales (see Line#6,7). ZoomNet improves upon the state-of-the-art methods by zooming in to the hand area for higher resolution. However, we also find some failure cases of our proposed ZoomNet. We observe that it still has difficulty in dealing with small face/hands with low-resolution and motion blur.

**Fig. 6.** Qualitative evaluation results of our approach in handling challenges including occlusion, close proximity, and small scale persons.


**Fig. 7.** Qualitative comparison between our proposed ZoomNet, OpenPose [1] and Single-Network [3]. Our approach outperforms the state-of-the-art approaches especially on face/hand keypoints and are more robust to scale variance.

## References


1. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: realtime multi-person 2d pose estimation using part affinity fields. arXiv preprint arXiv:1812.08008 (2018)
2. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
3. Hidalgo, G., Raaj, Y., Idrees, H., Xiang, D., Joo, H., Simon, T., Sheikh, Y.: Single-network whole-body pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
4. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
5. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proceedings of the European Conference on Computer Vision (ECCV) (2014)
6. Newell, A., Huang, Z., Deng, J.: Associative embedding: End-to-end learning for joint detection and grouping. In: Advances in Neural Information Processing Systems (2017)
7. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (NIPS) (2015)
8. Ronchi, M.R., Perona, P.: Benchmarking and Error Diagnosis in Multi-Instance Pose Estimation. Proceedings of International Conference on Computer Vision (ICCV) (2017)
9. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: IEEE International Conference on Computer Vision Workshop (2013)
10. Simon, T., Joo, H., Matthews, I., Sheikh, Y.: Hand keypoint detection in single images using multiview bootstrapping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
11. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. arXiv preprint arXiv:1902.09212 (2019)
12. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)