

# Single-Network Whole-Body Pose Estimation

Gines Hidalgo<sup>1</sup>, Yaadhav Raaj<sup>1</sup>, Haroon Idrees<sup>2</sup>, Donglai Xiang<sup>1</sup>, Hanbyul Joo<sup>3</sup>, Tomas Simon<sup>1</sup>, Yaser Sheikh<sup>1</sup>

<sup>1</sup>Carnegie Mellon University, <sup>2</sup>RetailNext, <sup>3</sup>Facebook AI Research

gines.hidalgo@epicgames.com, {ryaadhav,donglaix,yaser}@cs.cmu.edu, haroon@retailnext.net, {hjoo,tsimon}@fb.com



Figure 1: We present the first single-network approach for whole-body pose estimation, with real-time performance independent of the number of people in the image. Our work builds upon the current state-of-the-art OpenPose [9], boosting considerably its run-time performance while simultaneously improving slightly on the keypoint accuracy.

## Abstract

We present the first single-network approach for 2D whole-body pose estimation, which entails simultaneous localization of body, face, hands, and feet keypoints. Due to the bottom-up formulation, our method maintains constant real-time performance regardless of the number of people in the image. The network is trained in a single stage using multi-task learning, through an improved architecture which can handle scale differences between body/foot and face/hand keypoints. Our approach considerably improves upon OpenPose [9], the only work so far capable of whole-body pose estimation, both in terms of speed and global accuracy. Unlike [9], our method does not need to run an additional network for each hand and face candidate, making it substantially faster for multi-person scenarios. This work directly results in a reduction of computational complexity for applications that require 2D whole-body information (e.g., VR/AR, re-targeting). In addition, it yields higher accuracy, especially for occluded, blurry, and low resolution faces and hands. For code, trained models, and validation benchmarks, visit our project page<sup>1</sup>.

## 1. Introduction

Human keypoint estimation has been an open problem for decades in the research community. Initially, efforts were focused on facial alignment (i.e., face keypoint detection) [5, 52, 55, 72, 68, 74], and later evolved into single and multi-person human pose estimation in-the-wild, including body and foot keypoints [4, 7, 10, 13, 15]. A more recent and challenging problem has targeted hand keypoint detection [58, 61, 76]. Therefore, the next logical step is the integration of all of these keypoint detection tasks within the same algorithm, leading to “whole-body” or “full-body” (body, face, hand, and foot) pose estimation [1, 9].

There are several applications that can immediately take advantage of whole-body keypoint detection, including augmented reality, virtual reality, medical applications, and sports analytics. Whole-body keypoint detection can also provide more subtle cues for re-targeting, 3D human keypoint and mesh reconstruction [6, 11, 28, 42, 66], person re-identification, tracking, and action recognition [22, 35, 50, 49]. Despite these needs, the only existing method providing whole-body pose estimation is the prior version of OpenPose [9], which follows a multi-stage approach. First, all body poses are obtained from an input image in a bottom-up fashion [10] and then additional face and hand

<sup>1</sup>[https://github.com/CMU-Perceptual-Computing-Lab/openpose\\_train](https://github.com/CMU-Perceptual-Computing-Lab/openpose_train)

keypoint detectors are run for each detected person [58]. As a multi-network approach, it directly uses the existing body, face, and hand keypoint detection algorithms. However, it suffers from the issue of early commitment: there is no recourse to recovery if the body-only detector fails, especially during partial visibility when only the face or hand are visible in the image. In addition, its run-time is proportional to the number of people in the image, making whole-body pose estimation prohibitively expensive for multi-person and real-time applications. A single-stage method, estimating whole-body poses of multiple people in a single pass, would be more attractive as it would yield a fixed inference run-time, independent of the number of people in the scene.

Unfortunately, there is an inherent scale difference between body/foot and face/hand keypoints. The former require a large receptive field to learn the complex interactions across people (contact, occlusion, limb articulation), while the latter require higher image resolution. Since the foot pose is highly dependent on that of the body, unlike face and hands, its desirable scale is consistent with that of the body. Moreover, the scale issue has two critical consequences. First, datasets with full-body annotations in-the-wild do not currently exist, since the characteristics of each set of keypoints result in different kinds of datasets. Body datasets predominantly contain images with multiple people, usually resulting in fairly low face and hand resolution, while face and hand datasets mostly contain images with a single, cropped face or hand. Secondly, the architecture design of a single-network model must differ from that of the state-of-the-art keypoint detectors in order to offer high-resolution and a larger receptive field, while simultaneously improving the inference run-time of multi-network approaches.

To overcome the dataset problem, we resort to multi-task learning (MTL), a classic machine learning technique [19, 36, 73] where related learning tasks are solved simultaneously by exploiting commonalities and differences across them. Previously, MTL has been successful in training a combined body-foot keypoint detector [9]. Nevertheless, it does not generalize to whole-body estimation because of the underlying scale problem. Therefore, the major contributions of this paper are summarized as follows:

- **Novelty:** We present an MTL approach combined with an improved architecture design to train a unified model for various keypoint detection tasks each with different scale characteristics. This results in the *first* single-network approach for whole-body multi-person pose estimation.
- **Speed:** At test time, our single-network approach provides a constant real-time inference regardless of the number of people detected, and it is approximately  $n$  times faster than the state-of-the-art (OpenPose [9]) for images with  $n$  people. In addition, it is trained in a sin-

gle stage, rather than requiring independent network training for each individual task. This reduces the total training time approximately by one-half.

- **Accuracy:** Our approach also yields higher accuracy than that of the previous OpenPose, especially for face and hand keypoint detection, generalizing better to occluded, blurry, and low resolution faces and hands.

## 2. Related Work

**Face Keypoint Detection:** Also referred in literature as landmark detection or face alignment, it has a long history in computer vision and many approaches have been proposed to tackle it. These approaches are broadly divisible into two categories: template fitting [5, 31, 55, 68, 75] and regression-based methods [52, 72, 74]. Template fitting methods build face templates to fit input images, usually exploiting a cascade of regression functions. Regression methods, on the other hand, are based on Convolutional Neural Networks (CNNs) and usually apply convolutional heatmap regression. They operate in a similar fashion to that of body pose estimation.

**Body Keypoint Estimation:** With the face alignment problem solved, efforts moved towards single-person pose estimation. The initial approaches performed inference over both local observations on body parts and their spatial dependencies, either based on tree-structured graphical models [4, 18, 46, 51, 71] or non-tree models [15, 30, 33, 57, 63]. The popularity of CNNs and the release of massive annotated datasets (COCO [34] and MPII [2]) imparted significant boost in the accuracy of single-person estimation [7, 12, 14, 32, 39, 62, 64, 69], and have enabled multi-person estimation. The latter is traditionally divided into top-down [13, 17, 20, 23, 24, 44, 48, 67] and bottom-up [10, 38, 40, 43, 47] approaches.

**Foot Keypoint Estimation:** Cao *et al.* [9] released the first foot dataset, with annotations on a subset of images from the COCO dataset. They also trained the first combined body-foot keypoint detector by applying a naive multi-task learning technique. Our method is an extension of this work, mitigating its limitations and enabling it to generalize to both large-scale body and foot keypoints as well as the more subtle face and hand keypoints.

**Hand Keypoint Detection:** With the exciting improvements in face and body estimation, recent research is targeting hand keypoint detection. However, its manual annotation is extremely challenging and expensive due to heavy self-occlusion [58]. As a result, large hand keypoint datasets do not exist in-the-wild. To alleviate this problem, early work is based on depth information [41, 56, 60, 61], but is limited to indoor scenarios. Most of the work in RGB-based hand estimation is focused on 3D estimation [8, 25, 37, 76], primarily based on fitting complex 3D models with strong priors. In the 2D RGB domain, Si-

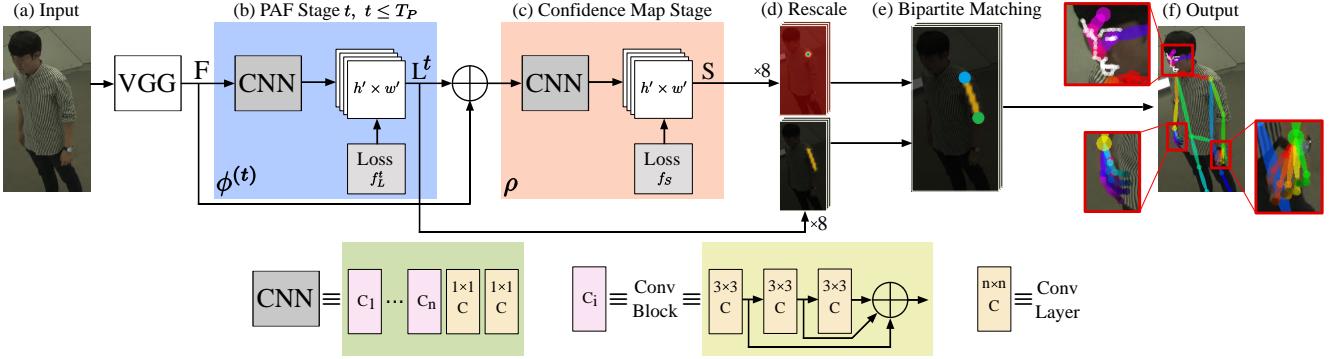


Figure 2: Overall pipeline. (a) An RGB image is taken as input. (b,c) Architecture of the whole-body pose estimation network, consisting of multiple stages predicting refined PAFs ( $L$ ) and confidence maps ( $S$ ) for body, face, hand and foot. It is trained end-to-end with a multi-task loss that combines the losses of each individual keypoint annotation task. Each *Conv Layer*,  $C$ , corresponds to a Convolution-PReLU sequence. (d) At test time, the most refined PAFs and confidence maps are resized to increase the accuracy. (e) The parsing algorithm uses the PAFs to find all the whole-body parts belonging to the same person using bipartite matching. (f) The final whole-body poses are returned for all the people in the image.

mon *et al.* [58] exploit multi-view bootstrapping to create a hand keypoint dataset and train a 2D RGB-based hand detector. First, a naive detector is trained on a small subset of manually labeled annotations. Next, this detector is applied in a 30-camera multi-view dome structure [26, 28] to obtain new annotations based on 3D reconstruction. Unfortunately, most of the methods have only demonstrated results in controlled lab environments.

**Whole-Body Keypoint Detection:** OpenPose [9, 10, 58] is the only known work able to provide all body, face, hand, and foot keypoints in 2D. It operates in a multi-network fashion. First, it detects the body and foot keypoints based on [10, 64]. Then, it approximates the face and hand bounding boxes based on the body keypoints, and applies a keypoint detection network for each subsequent face and hand candidate [58]. Recent work is also targeting 3D mesh reconstruction [28, 29], usually leveraging the lack of 3D datasets with the existing 2D datasets and detectors, or reconstructing the 3D surface of the human body from denser 2D human annotations [1].

**Multi-Task Learning:** To overcome the problems of state-of-the-art whole-body pose estimation, we aim to apply multi-task learning (MTL) to train a single whole-body estimation model out of the four different tasks: body, face, hand, and foot detection. MTL applied to deep learning can be split into soft and hard parameter sharing of hidden layers. In soft parameter sharing, each task has its own model, but the distance between the parameters is regularized to encourage them to be similar between models [16, 70]. Hard parameter sharing is the most commonly used MTL approach in computer vision, applied in many applications, such as facial alignment [73] or surface normal prediction [36]. Particularly, it has had a critical impact on object detection, where popular approaches such as Fast

R-CNN [19] exploit MTL to merge all previously independent object detection tasks into a single and improved detector. It considerably improved training and testing speeds as well as detection accuracy. Analogous to Fast R-CNN, our work brings together multiple and independent keypoint detection tasks into a unified framework. See [53] for a more detailed survey of multi-task learning literature.

### 3. Method

Our system follows a streamlined approach, using an RGB image to generate a set of whole-body human keypoints for each person detected. This global pipeline is illustrated in Fig. 2. The extracted keypoints contain information from the face, torso, arms, hands, legs, and feet. The network architecture of the proposed whole-body keypoint detector can be based on any state-of-the-art body-only keypoint detector. For a fair comparison of our results with previous versions of OpenPose [9, 10, 58], we reuse its Part Affinity Field (PAF) network architecture.

#### 3.1. PAF-based Body Pose Estimation

Here, we review the main details of the PAF-based method. We refer the reader to [10] for a full description. This approach iteratively predicts Part Affinity Fields (PAFs), which encode part-to-part associations, and detection confidence maps. Each PAF is defined as a 2D orientation vector that points from one keypoint to another. The input image  $I$  is initially analyzed by a convolutional network (pre-trained on VGG-19 [59]), generating a set of feature maps  $F$ . Next,  $F$  is fed into the first stage  $\phi^{(1)}$  of the network  $\phi$ , which predicts a set of PAFs  $L^{(1)}$ . For each subsequent stage  $i$ , the PAFs of the previous stage  $L^{(t-1)}$  are concatenated to  $F$  and refined to produce  $L^{(t)}$ . After  $N$

stages, we obtain the final set of PAF channels  $\mathbf{L} = \mathbf{L}^{(N)}$ . Then,  $\mathbf{F}$  and  $\mathbf{L}$  are concatenated and fed into a network  $\rho$ , which predicts the keypoint confidence maps  $\mathbf{S}$ , i.e.,

$$\mathbf{L}^{(1)} = \phi^{(1)}(\mathbf{F}), \quad (1)$$

$$\mathbf{L}^{(t)} = \phi^{(t)}\left(\mathbf{F}, \mathbf{L}^{(t-1)}\right), \forall 2 \leq t \leq N, \quad (2)$$

$$\mathbf{L} = \mathbf{L}^{(N)}, \quad (3)$$

$$\mathbf{S} = \rho(\mathbf{F}, \mathbf{L}). \quad (4)$$

An  $\ell_2$  loss function is applied at the end of each stage, which compares the estimated predictions and the groundtruth maps ( $\mathbf{S}^*$ ) and fields ( $\mathbf{L}^*$ ) for each pixel ( $p$ ) on each confidence map ( $c$ ) and PAF ( $f$ ) channel:

$$f_{\mathbf{L}} = \sum_{f=1}^F \sum_p (W_i(p) \cdot \|\mathbf{L}_f(p) - \mathbf{L}_f^*(p)\|_2^2), \quad (5)$$

$$f_{\mathbf{S}} = \sum_{c=1}^C \sum_p (W_i(p) \cdot \|\mathbf{S}_c(p) - \mathbf{S}_c^*(p)\|_2^2), \quad (6)$$

where  $C$  and  $F$  are the number of stages for confidence map and PAF prediction, and  $W$  is a binary mask with  $W_i(p)=0$  when an annotation is missing at a pixel  $p$  for a particular confidence map or PAF channel  $i$ . Non-maximum suppression is performed on the confidence maps to obtain a discrete set of body part candidate locations. Finally, bipartite graph matching [65] is used to assemble the connections that share the same part detection candidates into full-body poses for each person in the image.

### 3.2. Whole-Body Pose Estimation

We want whole-body pose estimation to be accurate but also fast. Training an individual PAF-based network to predict each individual set of keypoints would achieve the first goal, but would also be computationally inefficient. Instead, we extend the body-only PAF framework to whole-body pose estimation, making various modifications to the training approach and network architecture.

**Multi-task learning training:** We modify the definition of the keypoint confidence maps  $\mathbf{S}$  as the concatenation of the confidence maps: body  $S_B$ , face  $S_F$ , hand  $S_H$ , and foot  $S_O$ . Analogously, the set of PAFs at stage  $t$ ,  $\mathbf{L}^{(t)}$ , is defined as the concatenation of the PAFs: body  $\mathbf{L}_B^{(t)}$ , face  $\mathbf{L}_F^{(t)}$ , hand  $\mathbf{L}_H^{(t)}$ , and foot  $\mathbf{L}_O^{(t)}$ . An interconnection between the different annotation tasks must be created in order to allow the different set of keypoints of the same person to be assembled together. For instance, we join the body and foot keypoints through the ankle keypoint, which is annotated in both datasets. Analogously, the wrists connect the body and hand keypoints, while the eyes relate body and face. The rest of the pipeline (non-maximum suppression over confidence maps and bipartite matching to assemble



Figure 3: Different kinds of datasets for each set of keypoints present different properties (number of people, occlusion, person scale, etc.). We show typical examples from the hand (left), body (center), and face (right) datasets.

full people) is left intact. As opposed to having a dedicated network for each keypoint annotation task, all the keypoints are now defined within the same model architecture. This is an extreme version of hard parameter sharing, in which only the final layer is task-specific.

**Balanced dataset-based probability ratio:** If we had a whole-body dataset, we could train a combined model following the body-only training approach. Unfortunately, each available dataset only contains annotations for a subset of keypoints. To overcome the lack of a combined dataset, we follow the probability ratio idea of the single-network body-foot detector of Cao *et al.* [9], which was trained from body-only and body-foot datasets. Batches of images are randomly picked from each available dataset, and the losses for the confidence map and PAF channels associated to non-labeled keypoints are masked out. I.e., their binary mask  $W_i(p)$  is set to 0. Abusing notation, the probability ratio  $P^d$  is defined as the probability of picking the next annotated batch of images from the dataset  $d$ . This probability is distributed across the different datasets depending on the number of images in each dataset. When applied to keypoints with similar scale properties (e.g., body and foot [9]), it results in a robust keypoint detector. However, it does not converge when applied to whole-body estimation, which now includes face and hand keypoints. Additionally, the accuracy of the body and foot detectors is considerably reduced. Solving the face and hand convergence problem would be then possible through a deeper analysis of the properties and differences of each set of keypoints.

**Dataset-based augmentation:** There is an inherent scale difference between body/foot and face/hand keypoints, which results in different kinds of datasets for each set of keypoints. Body datasets predominately contain images with multiple people and low face and hand resolution; face datasets focus on images with a single person or cropped face; whereas hand datasets usually contain images with a single full-body person. Fig. 3 shows typical examples from each dataset. To solve this problem, augmentation parameters are varied for each set of keypoints. For instance, the minimum possible scale for face datasets

is reduced during data augmentation to expose our model to small faces, to simulate reality of the ‘wild’ environments. Oppositely, the maximum scale for hand datasets is increased so that full-sized hands appear more frequently, allowing the network to generalize to high resolution hands.

**Overfitting:** Using the above ideas, the face and hand detectors converge and allow us to build an initial whole-body pose detector. However, we observe a large degree of over-fitting on some validation sets, particularly in the face and lab-recorded datasets. Even though the initial probability ratio  $P^d$  is evenly distributed depending on the number of images in each dataset, the data complexity of these datasets is lower than the complexity of the challenging multi-person and in-the-wild datasets. In addition, the range of possible facial gestures is much smaller than the number of possible body and hand poses. Thus, the probability ratio of picking a batch from one of the face and lab-recorded datasets must be additionally reduced. Empirically, we fine-tune the probability ratios between datasets so that the validation accuracy on each converges at the same pace.

**High false positive rate:** Face, hand, and foot keypoints present a high false positive rate producing a “ghosting” effect on their respective confidence map and PAF channels. Visually, this means that these channels are outputting a non-zero value in regions that do not contain people. To mitigate this issue, their binary mask  $W_i(p)$  is re-enabled in the COCO dataset for image regions with no people. Furthermore, we complement training with an additional dataset consisting of COCO images without any people.

**Further refinement:** Face and hand datasets do not necessarily annotate all the people that appear in each image. We apply Mask R-CNN [23] to mask out the regions of the image with non-labeled people. In addition, the pixel localization precision of the face and hand keypoint detectors remains low. To moderately improve it, we reduce the radius of the Gaussian distribution used to generate the groundtruth of their confidence map channels.

**Shallow whole-body detector:** At this point, we can build a working whole-body pose detector. The inference run-time of this refined detector matches that of running body-foot in [9]. However, it still suffers from two main issues. On the one hand, the body and foot accuracy considerably decreases compared to its standalone analog (i.e., the body-foot detector from Cao *et al.* [9]). The complexity of the network output has increased from predicting 25 to 135 keypoints (and their corresponding PAFs). The network has to compress about 5 times more information with the same number of parameters, reducing the accuracy of each individual part. On the other hand, face and hand detection accuracy appear relatively similar to that of Cao *et al.* [9] in the benchmarks, but the qualitative results show that their pixel localization precision remains low. This is due to the reuse of same network as that used in body-only pose esti-

mation which has low input resolution. Face and hand detection requires a network with higher resolution to provide results with high pixel localization precision. This initial detector is defined as “Shallow whole-body” in Sec. 4.

**Improved network architecture:** To match the accuracy of the body-only detector and solve the resolution issue of face and hand, the whole-body network architecture must diverge from that of Cao *et al.* [9]. It must still maintain a large receptive field for accurate body detection but also offer high-resolution maps for precise face and hand keypoint detection. Additionally, its inference run-time should remain similar to or improve upon that of its analogous multi-stage whole-body detector. Our final model architecture, refined for whole-body estimation and shown in Fig. 2, differs from the original baseline in the following details:

- The network input resolution is increased to considerably improve face and hand precision. Unfortunately, this implicitly reduces the effective receptive field (further reducing body accuracy).
- The number of convolutional blocks in each PAF stage is increased to recover the effective receptive field that was previously reduced.
- The width of each convolutional layer in the last PAF stage is increased to improve the overall accuracy, enabling our model to match the body accuracy of the standalone body detector.
- The previous solutions considerably increase the overall accuracy of our approach but also harm the training and testing speed. The number of PAF stages is reduced to partially overcome this issue, which only results in a moderate reduction in overall accuracy.

This improved model highly outperforms Cao *et al.* [9] in speed, being approximately  $n$  times faster for an image with  $n$  people in it. Additionally, it also slightly improves its global accuracy (Secs. 4.3, 4.4 and 4.5). This network is denoted as “Deep whole-body” in Sec. 4.

## 4. Evaluation

### 4.1. Experimental Setup

**Datasets:** We train and evaluate our method on different benchmarks for each set of keypoints: (1) COCO keypoint dataset [34] for multi-person body estimation; (2) OpenPose foot dataset [9], which is a subset of 15k annotations out of the COCO keypoint dataset; (3) OpenPose hand dataset [58], which combines a subset of 1k hand instances manually annotated from MPII [2] as well as a set of 15k samples automatically annotated on the Dome or Panoptic Studio [27]; (4) our custom face dataset, consisting of a combination of the CMU Multi-PIE Face [21], Face Recognition Grand Challenge (FRGC) [45], and i-bug [54] datasets; (5) the Monocular Total Capture dataset [66],

the only available 2D whole-body dataset which has been recorded in the same Panoptic Studio used for the hand dataset. Following the standard COCO multi-person metrics, we report mean Average Precision (AP) and mean Average Recall (AR) for all sets of keypoints.

**Training:** All models are trained using 4-GPU machines, with a batch size of 10 images, Adam optimization, and an initial learning rate of 5e-5. We also decrease the learning rate by a factor of 2 after 200k, 300k, and every additional 60k iterations. We apply random cropping, rotation ( $\pm 45^\circ$ ), flipping (50%), and scale (in the range [1/3, 1.5]) augmentation. The scale is modified to [2/3, 4.5] and [0.5, 4.0] for Dome and MPII hand datasets, respectively. The input resolution of the network is set to  $480 \times 480$  pixels. Similar to [9], we maintain VGG-19 as the backbone. The probability of picking an image from each dataset is 76.5% for COCO, 5% each for foot and MPII datasets, 0.33% for each face dataset, 0.5% for Dome hand, 5% for MPII hand, 5% for whole-body data, and 2% for picking an image with no people in it.

**Evaluation:** We report both single-scale (image resized to a height of 480 pixels while maintaining the aspect ratio) and multi-scale results (results averaged from images resized to a height of 960, 720, 480, and 240 pixels).

## 4.2. Ablation Experiments

Increasing the network resolution is crucial for accurate hand and face detection. Nevertheless, it directly results in slower training and testing speeds. We aim to maximize the accuracy while preserving a reasonable run-time performance. Thus, we explore multiple models tuned to maintain the same inference run-time. The final model is selected as the one maximizing the body AP. Table 1 shows the results on the COCO [34] validation set. The most efficient configuration is achieved by increasing the number of convolutional blocks and their width, while reducing the number of stages in order to preserve the speed.

## 4.3. Body and Foot Keypoint Detection Accuracy

Once the optimal model has been selected, it is trained for whole-body estimation. Table 2 show the accuracy results on the COCO validation set for our 3 different models, as well as the results reported by Cao *et al.* [9]. The new deeper architecture slightly increases the accuracy of OpenPose when trained for whole-body estimation. It can also be applied to body-foot estimation, achieving a 1.1% improvement in accuracy compared to that of Cao *et al.* [9]. Interestingly, adding face and hand keypoints to the same model results in a considerable decrease of the body detection accuracy by about 5% for the shallow model when compared to that of [9]. Intuitively, this is due to the fact that we are trying to fit nearly six times as many keypoints into the same network. The original model might not have

Method		AP	AR	APs	ARs
PAF	CM				
1s, 10b, 256w	1s, 10b, 256w	65.8	70.3	56.1	61.1
2s, 8b, 128-288w	1s, 8b, 256w	66.1	70.5	56.7	61.9
2s, 10b, 128-256w	1s, 10b, 256w	66.1	70.7	<b>57.0</b>	<b>62.0</b>
3s, 8b, 96-256w	1s, 8b, 192w	<b>66.4</b>	<b>70.9</b>	56.9	61.9
4s, 8b, 96-256w	1s, 8b, 224w	65.7	70.2	56.3	61.4
5s, 8b, 64-256w	1s, 5b, 256w	65.5	70.1	56.7	61.8

Table 1: Self-comparison on the body COCO validation set. All models have been tuned to have the same inference run-time. ‘APs’ and ‘ARs’ refer to the single-scale results. ‘PAF’ represents the Part Affinity Field network configuration and ‘CM’ the confidence map configuration. ‘s’ refers to the number of stages of refinement, ‘b’ to the number of convolutional blocks per stage, ‘w’ to the number of output channels (or width) of each convolutional layer. All other settings follow Sec. 4.1.

enough capacity to handle the additional complexity introduced by the new keypoints. However, this gap is smaller than 1% for the improved architecture (deep body-foot vs. deep whole-body). The additional depth helps the network generalize to a higher number of output keypoints.

Method	Body AP	Foot AP
Body-foot OpenPose (multi-scale) [9]	65.3	<b>77.9</b>
Shallow whole-body (ours, multi-scale)	60.9	70.2
Deep body-foot (ours, multi-scale)	<b>66.4</b>	76.8
Deep whole-body (ours, multi-scale)	65.6	76.2

Table 2: Accuracy results on the COCO validation set. ‘Shallow’ refers to the network architecture with the same depth and input resolution as that of OpenPose, while ‘Deep’ refers to our improved architecture. ‘Body-foot’ refers to the network that simply predicts body and foot keypoints, following the default OpenPose output, while ‘Whole-body’ refers to our novel single-network model.

## 4.4. Face Keypoint Detection Accuracy

In order to evaluate the accuracy of face alignment, traditional approaches have used the Probability of Correct Keypoint (PCK) metric, which checks the probability that a predicted keypoint is within a distance threshold of its true location. However, it does not generalize to a multi-person setting. In order to evaluate our work, we reuse the mean Average Precision (AP) and Recall (AR), following the COCO multi-person metric. We train our whole-body algorithm with the same face datasets that OpenPose [9] used: Multi-PIE [21], FRGC [45], and i-bug [54]. We create a custom validation set by selecting a small subset of images from each dataset. We show the results in Table 3. We can see that both our method and [9] greatly over-fit on

the Multi-PIE and FRGC datasets. These datasets consist of images annotated in controlled lab environments, and all faces appear frontal and with no occlusion, similar to the last image in Fig. 3. However, their accuracy is considerably lower in the in-the-wild i-bug dataset, where our approach is about 2% more accurate.

Method	Face AR		
	FRGC	M-Pie	i-bug
OpenPose [9]	98.3	<b>96.3</b>	52.4
Shallow Whole-body (ours, single-scale)	<b>98.4</b>	90.6	50.6
Deep Whole-body (ours, single-scale)	<b>98.4</b>	93.2	<b>54.5</b>

Table 3: Accuracy results on our custom CMU Multi-PIE and FRGC validation sets. All the people in each image are not necessarily labeled on i-bug. Thus, those samples might be considered erroneous “false positives” and affect the AP results. However, AR is only affected by the annotated samples, so it is used as the main metric for i-bug.

#### 4.5. Hand Keypoint Detection Accuracy

Analog to face evaluation, we randomly select a subset of images from each hand dataset for validation. We denote “Hand Dome” for the subset of [58] recorded in the Panoptic Studio [27], and “Hand MPII” for the subset manually annotated from MPII [3] images. The results are presented in Table 4. Both our method and the previous OpenPose over-fit on the Dome dataset, where usually only a single person appears in each frame, similar to the first image in Fig. 3. However, the manually annotated images from MPII are more challenging for both approaches, as it represents realistic in-the-wild scenes. In such images, we can see the clear benefit of our deeper network with respect to Cao *et al.* [9] and our initial shallow model, outperforming by about 5.5% on the Hand MPII dataset.

#### 4.6. Run-time Comparison

In Fig. 4, we compare the inference run-time between OpenPose and our work. Our method is 10% faster than OpenPose for images with a single person. However, the inference time of our single-network approach remains constant, while OpenPose’s time is proportional to the number

Method	Hand AR	
	Dome	MPII
OpenPose (single-scale) [9]	97.0	82.7
Shallow whole-body (ours, single-scale)	94.6	82.4
Deep whole-body (ours, single-scale)	<b>97.8</b>	<b>88.1</b>

Table 4: Results on our custom Hand Dome and Hand MPII validation sets. These datasets might contain unlabeled people (similar to i-bug), so AR is used for evaluation.

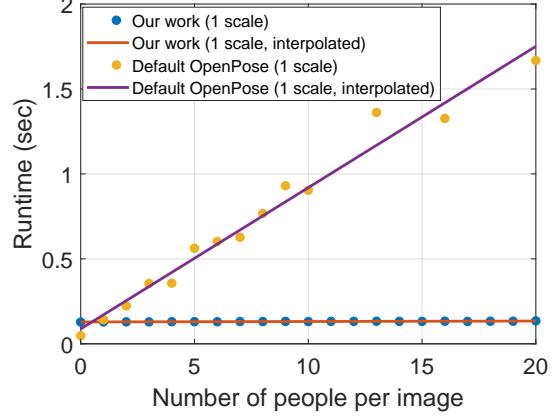


Figure 4: Inference time comparison between our work and whole-body OpenPose [9]. While the inference time of our proposed approach is invariant, the run-time of OpenPose grows linearly with the number of people. OpenPose runtime presents some oscillations because it does not run face and hand detectors if the nose or wrist keypoints (provided by the body network) of a person are not found. This is a common case in images with many crowded images. This analysis was performed on a system with a Nvidia 1080 Ti.

of people detected, in particular, it is proportional to the number of face and hand proposals. This leads to a massive speedup of our approach when the number of people increases. For images with  $n$  people, our approach is approximately  $n$  times faster than OpenPose. For crowded images, many hands and faces are occluded, slightly reducing this speedup. For instance, our approach is about 7 times faster than OpenPose for typical images with 10 people in them.

## 5. Conclusion

In this paper, we resort to multi-task learning combined with an improved model architecture to train the first single-network approach for 2D whole-body estimation. Our work brings together multiple and, currently, independent keypoint detection tasks into a unified framework. We evaluate our method on multiple keypoint detection benchmarks and compare it with the state-of-the-art, considerably outperforming it in both training and testing speed as well as slightly improving its accuracy. We qualitatively show in Fig. 5 that our face and hand detectors generalize better to in-the-wild images, benefiting from their indirect exposure to the immense body datasets. Nevertheless, there are still some limitations with our method. First, we observe global failure cases when a significant part of the target person is occluded or outside of the image boundaries. Secondly, the accuracy of the face and especially hand keypoint detectors is still limited, failing in the case of severe motion blur, small people, and extreme gestures. Third, we qualitatively observe that the previous version of OpenPose outperforms

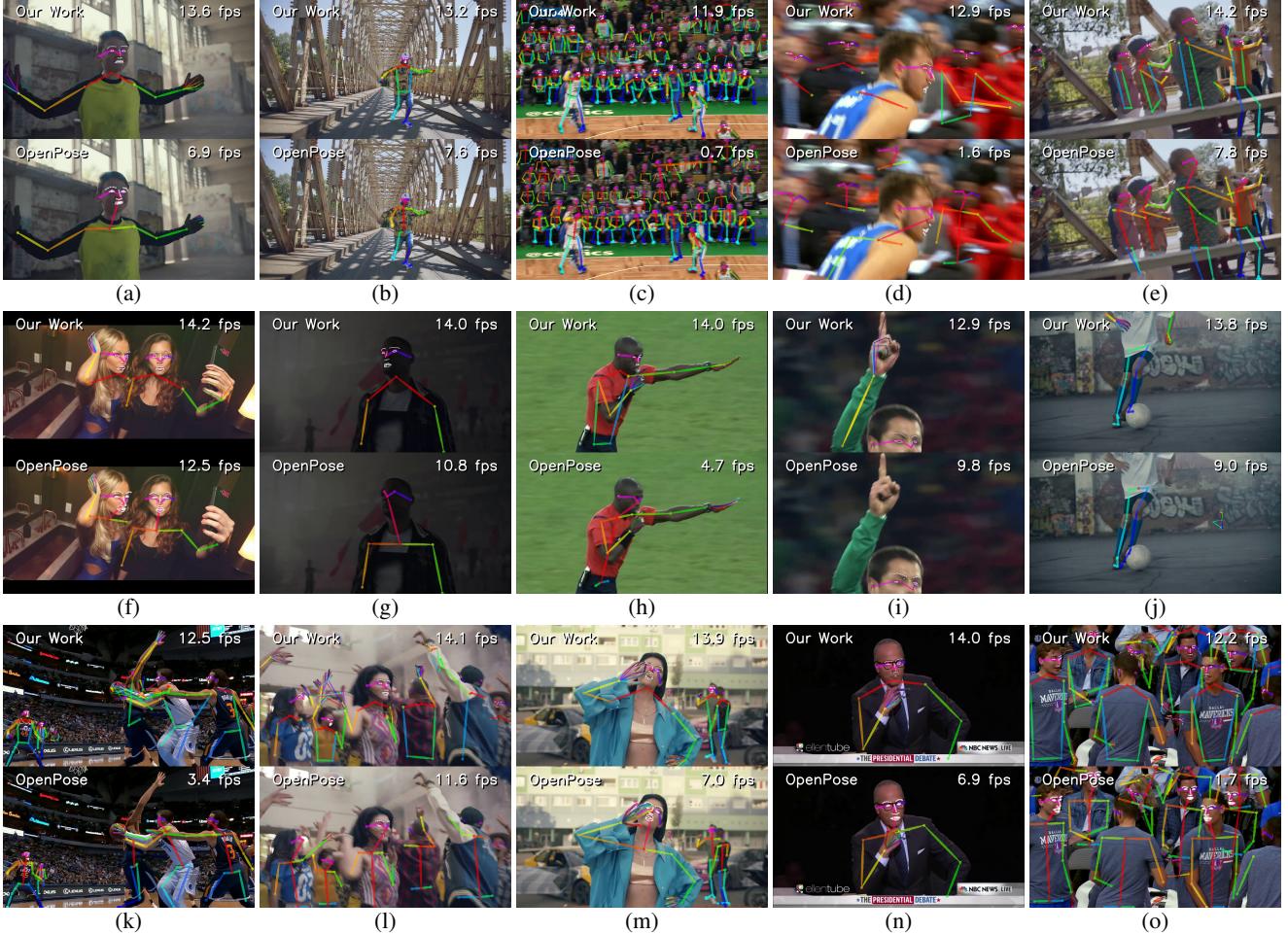


Figure 5: Qualitative comparison between our method (top) and the previous version of OpenPose [9] (bottom), performed on a system with 2 Nvidia 1080 Ti. (a-j) show improved results and (k-o) failing cases. (a) It generalizes better to hands wearing any kind of gloves. (b) The implicit finger information helps wrist and elbow detection. (c) Even smaller faces and hands are detected. (d-e) Blurry and profile faces are detected. (f) More extreme hand poses are detected. (g) Faces from low-brightness images are better detected. (h) Hands where all fingers are occluded are detected. (i-j) Cropped arms are properly detected. (k-l) It shows difficulties when several hands are in proximity. (m-o) It seems to fail for some relatively easy hand and face poses that are successfully detected with the previous version of OpenPose.

ours for face and hand detection when poses are simple and no occlusion occurs. The previous method crops the bounding box proposal of those bounding box candidates, resizes them up, and feeds them into its dedicated networks. This higher input resolution leads to an increased pixel localization precision if the keypoint detection is successful.

## References

- [1] Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, pages 7297–7306, 2018. [1](#), [3](#)
- [2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: new benchmark and state of the art analysis. In *CVPR*, 2014. [2](#), [5](#)
- [3] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, June 2014. [7](#)
- [4] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Monocular 3D pose estimation and tracking by detection. In *CVPR*, 2010. [1](#), [2](#)
- [5] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. Incremental face alignment in the wild. In *CVPR*, pages 1859–1866, 2014. [1](#), [2](#)
- [6] Aayush Bansal, Shugao Ma, Deva Ramanan, and Yaser Sheikh. Recycle-gan: Unsupervised video retargeting. In *ECCV*, 2018. [1](#)
- [7] Vasileios Belagiannis and Andrew Zisserman. Recurrent human pose estimation. In *IEEE FG*, 2017. [1](#), [2](#)

- [8] Yujun Cai, Liuhan Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *ECCV*, pages 666–682, 2018. 2
- [9] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*, 2018. 1, 2, 3, 4, 5, 6, 7, 8
- [10] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 1, 2, 3
- [11] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *ECCV Workshop*, 2018. 1
- [12] Yu Chen, Chunhua Shen, Xiu-Shen Wei, Lingqiao Liu, and Jian Yang. Adversarial posenet: A structure-aware convolutional network for human pose estimation. In *ICCV*, 2017. 2
- [13] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *CVPR*, 2018. 1, 2
- [14] Xiao Chu, Wei Yang, Wanli Ouyang, Cheng Ma, Alan L Yuille, and Xiaogang Wang. Multi-context attention for human pose estimation. In *CVPR*, 2017. 2
- [15] Matthias Dantone, Juergen Gall, Christian Leistner, and Luc Van Gool. Human pose estimation using body parts dependent joint regressors. In *CVPR*, 2013. 1, 2
- [16] Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *ACL-IJCNLP*, volume 2, pages 845–850, 2015. 3
- [17] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017. 2
- [18] Pedro F Felzenszwalb and Daniel P Huttenlocher. Pictorial structures for object recognition. In *IJCV*, 2005. 2
- [19] Ross Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015. 2, 3
- [20] Georgia Gkioxari, Bharath Hariharan, Ross Girshick, and Jitendra Malik. Using k-poselets for detecting people and localizing their keypoints. In *CVPR*, 2014. 2
- [21] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010. 5, 6
- [22] Liangyan Gui, Kevin Zhang, Yu-Xiong Wang, Xiaodan Liang, José MF Moura, and Manuela M Veloso. Teaching robots to predict human motion. In *IROS*, 2018. 1
- [23] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 2, 5
- [24] Umar Iqbal and Juergen Gall. Multi-person pose estimation with local joint-to-person associations. In *ECCV Workshop*, 2016. 2
- [25] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *ECCV*, pages 118–134, 2018. 2
- [26] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *ICCV*, 2015. 3
- [27] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE TPAMI*, 2017. 5, 7
- [28] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *CVPR*, 2018. 1, 3
- [29] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, pages 7122–7131, 2018. 3
- [30] Leonid Karlinsky and Shimon Ullman. Using linking features in learning non-parametric part models. In *ECCV*, 2012. 2
- [31] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *CVPR*, pages 1867–1874, 2014. 2
- [32] Lipeng Ke, Ming-Ching Chang, Honggang Qi, and Siwei Lyu. Multi-scale structure-aware network for human pose estimation. In *ECCV*, 2018. 2
- [33] Xiangyang Lan and Daniel P Huttenlocher. Beyond trees: Common-factor models for 2d human pose recovery. In *ICCV*, 2005. 2
- [34] 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. 2, 5, 6
- [35] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. In *ACM TOG*, 2017. 1
- [36] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-stitch networks for multi-task learning. In *CVPR*, pages 3994–4003, 2016. 2, 3
- [37] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *CVPR*, pages 49–59, 2018. 2
- [38] Alejandro Newell, Zhao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *NIPS*, 2017. 2
- [39] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 2
- [40] Xuecheng Nie, Jiashi Feng, Junliang Xing, and Shuicheng Yan. Pose partition networks for multi-person pose estimation. In *ECCV*, 2018. 2
- [41] Iasonas Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Tracking the articulated motion of two strongly interacting hands. In *CVPR*, pages 1862–1869. IEEE, 2012. 2
- [42] Paschalis Panteleris, Iason Oikonomidis, and Antonis Argyros. Using a single rgb frame for real time 3d hand pose estimation in the wild. In *IEEE WACV*, pages 436–445. IEEE, 2018. 1

- [43] George Papandreou, Tyler Zhu, Liang-Chieh Chen, Spyros Gidaris, Jonathan Tompson, and Kevin Murphy. Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In *ECCV*, 2018. 2
- [44] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards accurate multi-person pose estimation in the wild. In *CVPR*, 2017. 2
- [45] P Jonathon Phillips, Patrick J Flynn, Todd Scruggs, Kevin W Bowyer, Jin Chang, Kevin Hoffman, Joe Marques, Jaesik Min, and William Worek. Overview of the face recognition grand challenge. In *CVPR*, volume 1, pages 947–954. IEEE, 2005. 5, 6
- [46] Leonid Pishchulin, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. Poselet conditioned pictorial structures. In *CVPR*, 2013. 2
- [47] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Björn Andres, Mykhaylo Andriluka, Peter Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *CVPR*, 2016. 2
- [48] Leonid Pishchulin, Arjun Jain, Mykhaylo Andriluka, Thorsten Thormählen, and Bernt Schiele. Articulated people detection and pose estimation: Reshaping the future. In *CVPR*, 2012. 2
- [49] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. Pose-normalized image generation for person re-identification. In *ECCV*, 2018. 1
- [50] Yaadhav Raaj, Haroon Idrees, Gines Hidalgo, and Yaser Sheikh. Efficient online multi-person 2d pose tracking with recurrent spatio-temporal affinity fields. In *CVPR*, 2019. 1
- [51] Deva Ramanan, David A Forsyth, and Andrew Zisserman. Strike a Pose: Tracking people by finding stylized poses. In *CVPR*, 2005. 2
- [52] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. Hyphereface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE TPAMI*, 41(1):121–135, 2019. 1, 2
- [53] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017. 3
- [54] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: Database and results. *Image and vision computing*, 47:3–18, 2016. 5, 6
- [55] Jason M Saragih, Simon Lucey, and Jeffrey F Cohn. Face alignment through subspace constrained mean-shifts. In *ICCV*, pages 1034–1041. IEEE, 2009. 1, 2
- [56] Toby Sharp, Cem Keskin, Duncan Robertson, Jonathan Taylor, Jamie Shotton, David Kim, Christoph Rhemann, Ido Leichter, Alon Vinnikov, Yichen Wei, et al. Accurate, robust, and flexible real-time hand tracking. In *ACM CHI*, pages 3633–3642. ACM, 2015. 2
- [57] Leonid Sigal and Michael J Black. Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In *CVPR*, 2006. 2
- [58] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017. 1, 2, 3, 5, 7
- [59] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3
- [60] Srinath Sridhar, Franziska Mueller, Antti Oulasvirta, and Christian Theobalt. Fast and robust hand tracking using detection-guided optimization. In *CVPR*, pages 3213–3221, 2015. 2
- [61] Srinath Sridhar, Antti Oulasvirta, and Christian Theobalt. Interactive markerless articulated hand motion tracking using rgb and depth data. In *ICCV*, pages 2456–2463, 2013. 1, 2
- [62] Wei Tang, Pei Yu, and Ying Wu. Deeply learned compositional models for human pose estimation. In *ECCV*, 2018. 2
- [63] Yang Wang and Greg Mori. Multiple tree models for occlusion and spatial constraints in human pose estimation. In *ECCV*, 2008. 2
- [64] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional Pose Machines. In *CVPR*, 2016. 2, 3
- [65] Douglas B. West et al. *Introduction to graph theory*, volume 2. Prentice hall Upper Saddle River, NJ, 1996. 4
- [66] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *CVPR*, pages 10965–10974, 2019. 1, 5
- [67] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018. 2
- [68] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, pages 532–539, 2013. 1, 2
- [69] Wei Yang, Shuang Li, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Learning feature pyramids for human pose estimation. In *ICCV*, 2017. 2
- [70] Yongxin Yang and Timothy M Hospedales. Trace norm regularised deep multi-task learning. *arXiv preprint arXiv:1606.04038*, 2016. 3
- [71] Yi Yang and Deva Ramanan. Articulated human detection with flexible mixtures of parts. In *IEEE TPAMI*, 2013. 2
- [72] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE SPL*, 23(10):1499–1503, 2016. 1, 2
- [73] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, pages 94–108. Springer, 2014. 2, 3
- [74] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning deep representation for face alignment with auxiliary attributes. *IEEE TPAMI*, 38(5):918–930, 2016. 1, 2
- [75] Shizhan Zhu, Cheng Li, Chen Change Loy, and Xiaoou Tang. Face alignment by coarse-to-fine shape searching. In *CVPR*, pages 4998–5006, 2015. 2
- [76] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *ICCV*, pages 4903–4911, 2017. 1, 2