

Monocular Real-time Full Body Capture with Inter-part Correlations

Yuxiao Zhou¹ Marc Habermann^{2,3} Ikhsanul Habibie^{2,3} Ayush Tewari^{2,3} Christian Theobalt^{2,3} Feng Xu^{1*}

¹BNRist and School of Software, Tsinghua University ²Max Planck Institute for Informatics ³Saarland Informatics Campus

Abstract

We present the first method for real-time full body capture that estimates shape and motion of body and hands together with a dynamic 3D face model from a single color image. Our approach uses a new neural network architecture that exploits correlations between body and hands at high computational efficiency. Unlike previous works, our approach is jointly trained on multiple datasets focusing on hand, body or face separately, without requiring data where all the parts are annotated at the same time, which is much more difficult to create at sufficient variety. The possibility of such multi-dataset training enables superior generalization ability. In contrast to earlier monocular full body methods, our approach captures more expressive 3D face geometry and color by estimating the shape, expression, albedo and illumination parameters of a statistical face model. Our method achieves competitive accuracy on public benchmarks, while being significantly faster and providing more complete face reconstructions.

1. Introduction

Human motion capture from a single color image is an important and widely studied topic in computer vision. Most solutions are unable to capture local motions of hands and faces together with full body motions. This renders them unsuitable for a variety of applications, e.g. AR, VR, or tele-presence, where capturing full human body pose and shape, including hands and face, is highly important. In these applications, monocular approaches should ideally recover the full body pose (including facial expression) as well as a render-ready dense surface which contains person-specific information, such as facial identity and body shape. Moreover, they should run at real-time framerates. Much progress has been made on relevant subtasks, i.e. body pose estimation [33, 31, 45, 40], hand pose estimation [78, 42, 80], and face capture [14, 61, 60, 53, 81]. How-

*This work was supported by the National Key R&D Program of China 2018YFA0704000, the NSFC (No.61822111, 61727808), Beijing Natural Science Foundation (JQ19015), and the ERC Consolidator Grant 4DRepLy (770784). Feng Xu is the corresponding author.



Figure 1: We present the first real-time monocular approach that jointly captures shape and pose of body and hands together with facial geometry and color. Top: results on in-the-wild sequences. Bottom: real-time demo. Our approach predicts facial color while the body color is set manually.

ever, joint *full body* capture, let alone in real-time, is still an open problem. Several recent works [9, 68, 28, 46, 38] have demonstrated promising results on capturing the full body. Nevertheless, they either only recover sparse 2D keypoints [38, 28], require specific training data [9, 28] where body, hands, and face are annotated altogether which is expensive to collect, or cannot achieve real-time performance [9, 68, 46, 38].

We therefore introduce the first real-time monocular approach that estimates: 1) 2D and 3D keypoint positions of body and hands; 2) 3D joint angles and shape parameters of body and hands; and 3) shape, expression, albedo, and illumination parameters of a 3D morphable face model [61, 14]. To recover the dense mesh, we use the SMPLH model [49] for body and hands surface, and replace its face area with a more expressive face model.

To achieve real-time performance without the loss of accuracy, we rigorously design our new network architecture to exploit inter-part correlations by streaming body features into the hand pose estimation branch. Specifically, the sub-network for hand keypoint detection takes in two sources

of features: one comes from the body keypoint detection branch as low-frequency global features, whereas the other is extracted from the hand area in the input image as high-frequency local features. This feature composition utilizes body information for hand keypoint detection, and saves the computation of extracting high-level features for the hands, resulting in reduced runtime and improved accuracy.

Further, we do not require a dataset where ground truth body, hands, and face reconstructions are all available at the same time: creating such data at sufficient variety is very difficult. Instead, we only require existing part-specific datasets. Our network features four task-specific modules that are trained individually with different types of data, while being end-to-end at inference. The first module, *DetNet*, takes a color image as input, estimates 3D body and hand keypoint coordinates, and detects the face location in the input image. The second and third module, namely *BodyIKNet* and *HandIKNet*, take in body and hand keypoint positions and regress joint rotations along with shape parameters. The last module, called *FaceNet*, takes in a face image and predicts the shape, expression, albedo, and illumination parameters of the 3DMM face model [61]. This modular network design enables us to jointly use the following data types: 1) images with only body *or* hand keypoint annotations; 2) images with body *and* hand keypoint annotations; 3) images annotated with body joint angles; 4) motion capture (MoCap) data with only body *or* hand joint angles but without corresponding images; and 5) face images with 2D landmarks. To train with so many data modalities, we propose an *attention* mechanism to handle various data types in the same mini-batch during training, which guides the model to utilize the features selectively. We also introduce a 2-stage body keypoint detection structure to cope with the keypoint discrepancy between different datasets. The above multi-modal training enables our superior generalization across different benchmarks.

Our contribution can be summarized as follows:

- The first real-time approach that jointly captures 3D body, hands and face from a single color image.
- A novel network structure that combines local and global features and exploits inter-part correlations for hand keypoint detection, resulting in high computational efficiency and improved accuracy.
- The utilization of various data modalities supported by decoupled modules, an attention mechanism, and a 2-stage body keypoint detection structure, resulting in superior generalization.

2. Related Work

Human performance capture has a long research history. Some methods are based on multi-view systems or a monocular depth camera to capture body [75, 29], hand [71, 43], and face [20, 50]. Although accurate, they

are largely limited by the hardware requirements: multi-view systems are hard to setup while depth sensors do not work under bright sunlight. This can be avoided by using a single RGB camera. As our approach falls in the category of monocular methods, we focus on related works that only require a monocular image.

Body and Hand Capture. The very early researches [55, 12] propose to combine local features and spatial relationship between body parts for pose estimation. With the advent of deep learning, new breakthrough is being made, from 2D keypoint detection [8, 15] to 3D keypoint estimation [58, 24, 39, 3]. In addition to sparse landmarks, recent approaches stress the task of producing a dense surface. A series of statistical parametric models [2, 36, 46, 30] are introduced and many approaches are proposed to estimate joint rotations for mesh animation. Some of these work [40, 54, 68] incorporate a separate inverse kinematics step to solve for joint rotations, while others [31, 33, 23] regress model parameters from input directly. To cope with the lack of detail in parametric models, some methods [69, 22, 23] propose to use subject-specific mesh templates and perform dense tracking of the surface with non-rigid deformations. Apart from model-based methods, model-free approaches also achieve impressive quality. Various surface representations are proposed, including mesh [34], per-pixel depth [17] and normal [57], voxels [76, 27], and implicit surface functions [51, 52]. The research of hand capture has a similar history. The task evolves from 2D keypoint detection [56, 65], to 3D keypoint estimation [79, 42, 13], and finally dense surface recovery [7, 78, 74, 72] based on parametric models [49, 63]. Methods that directly regresses mesh vertices are also proposed [41, 19, 4]. However, they all focus only on body or hands and failed to capture them jointly.

Face Capture. Early works [48, 18, 62, 66] reconstruct faces based on iterative optimization. Deep learning approaches [47, 64] are also presented in the literature. To cope with the problem of limited training data, semi- and self-supervised approaches are introduced [61, 60, 53, 59], where the models are trained in an analysis-by-synthesis fashion using differentiable rendering. We refer to the surveys [81, 14] for more details.

Full Body Capture. Several recent works investigate the task of capturing body, face and hands simultaneously from a monocular color image. The work of [67] estimates 3D keypoints of full body by distilling knowledge from part experts. To obtain joint angles, previous works [68, 46] propose a two-stage approach that first uses a network to extract keypoint information and then fits a body model onto the keypoints. Choutas et al. [9] regress model parameters directly from the input image and then apply hand/face-specific models to refine the capture iteratively. Although they demonstrate promising results, they are all far from be-

ing real-time. The shared shortcoming of their approaches is that they do not consider the correlation between body and hands. In their work, body information is merely used to locate [68, 9, 46] and initialize [9] hands, while we argue that the high-level body features can help to deduce the hand pose [44]. Further, recent methods [68, 46, 9] only capture facial expression, while our approach also recovers the facial identity in terms of geometry and color.

3. Method

As shown in Fig. 2, our method takes a color image as input, and outputs 2D and 3D keypoint positions, joint angles, and shape parameters of body and hands, together with facial expression, shape, albedo, and illumination parameters. We then animate our new parametric model (Sec. 3.1) to recover a dense full body surface. To leverage various data modalities, the whole network is trained as four individual modules: *DetNet* (Sec. 3.2) that estimates body and hand keypoint positions from a body image, with our novel inter-part feature composition, the attention mechanism, and the 2-stage body keypoint detection structure; *BodyIKNet* and *HandIKNet* (Sec. 3.3) that estimate shape parameters and joint angles from keypoint coordinates for body and hands; and *FaceNet* (Sec. 3.4) that regresses face parameters from a face image crop.

3.1. Full Body Model

Body with Hands. We use the SMPLH-neutral [49] model to represent the body and hands. Specifically, SMPLH is formulated as

$$T_B = \bar{T}_B + \beta E_\beta \quad (1)$$

where \bar{T}_B is the mean body shape with $N_B = 6890$ vertices, E_β is the PCA basis accounting for different body shapes, and values in $\beta \in \mathbb{R}^{16}$ indicate PCA coefficients. Given the body pose θ_b and the hand pose θ_h , which represent the rotation of $J_B = 22$ body joints and $J_H = 15 \times 2$ hand joints, the posed mesh is defined as

$$V_B = W(T_B, \mathcal{W}, \theta_b, \theta_h) \quad (2)$$

where $W(\cdot)$ is the linear blend skinning function and \mathcal{W} are the skinning weights.

Face. For face capture, we adopt the 3DMM [5] face model used in [61]. Its geometry is given as

$$V_F = \bar{V}_F + \zeta E_\zeta + \epsilon E_\epsilon \quad (3)$$

where \bar{V}_F is the mean face with $N_F = 53490$ vertices, E_ζ and E_ϵ are PCA bases that encode shape and expression variations, respectively. $\zeta \in \mathbb{R}^{80}$ and $\epsilon \in \mathbb{R}^{64}$ are the shape and expression parameters to be estimated. The face color is given by

$$R = \bar{R} + \gamma E_\gamma \quad (4)$$

$$t_i = r_i \sum_{b=1}^{B^2} \mu_b H_b(n_i) \quad (5)$$

where R and r_i are per vertex reflection, \bar{R} is the mean skin reflectance, E_γ is the PCA basis for reflectance, t_i and n_i are radiosity and normal of vertex i , and $H_b : \mathbb{R}^3 \rightarrow \mathbb{R}$ are the spherical harmonics basis functions. We set $B^2 = 9$. $\gamma \in \mathbb{R}^{80}$ and $\mu \in \mathbb{R}^{3 \times 9}$ are albedo and illumination parameters.

Combining Face and Body. To replace the SMPLH face with the 3DMM face, we manually annotate the face boundary \mathcal{B}_b of SMPLH and the corresponding boundary \mathcal{B}_f on the 3DMM face. Then, a rigid transformation with a scale factor is manually set to align the face-excluded part of \mathcal{B}_b and the face part of \mathcal{B}_f . This manual work only needs to be performed once. After bridging the two boundaries using Blender [11], the face part rotates rigidly by the upper-neck joint using the head angles. Unlike previous works [46, 30], we do not simplify the face mesh. Our model has more face vertices ($N'_F = 23817$) than the full body meshes of [9, 46] (10475 vertices) and [30, 68] (18540 vertices), supports more expression parameters (64 versus 40 [30, 68] and 10 [9, 46]), and embeds identity and color variation for face while others do not. This design allows us to model face more accurately and account for the fact that humans are more sensitive to the face quality. We show the combination process and full body meshes in Fig. 3.

3.2. Keypoint Detection Network: DetNet

The goal of our keypoint detection network, *DetNet*, is to estimate 3D body and hand keypoint coordinates from the input image. Particularly challenging is that body and hands have very different scales in an image so that a single network can barely deal with both tasks at the same time. The naive solution would be to use two separate networks. However, they would require much longer runtime, making real-time difficult to achieve. Our key observation to solve this issue is that the high-level global features of the hand area extracted by the body keypoint estimation branch can be shared with the hand branch. By combining them with the high-frequency local features additionally extracted from the hand area, expensive computation of hand high-level features is avoided, and body information for hand keypoint detection is provided, resulting in higher accuracy.

3.2.1 Two-Stage Body Keypoint Detection

It is a well-known issue that different body datasets have different sets of keypoint definitions, and the same keypoint is annotated differently in different datasets [30]. This inconsistency prevents the utilization of multiple datasets to improve the generalization ability. To this end, instead of estimating all keypoints at once, we follow a two-stage manner for body keypoint detection. We split the body

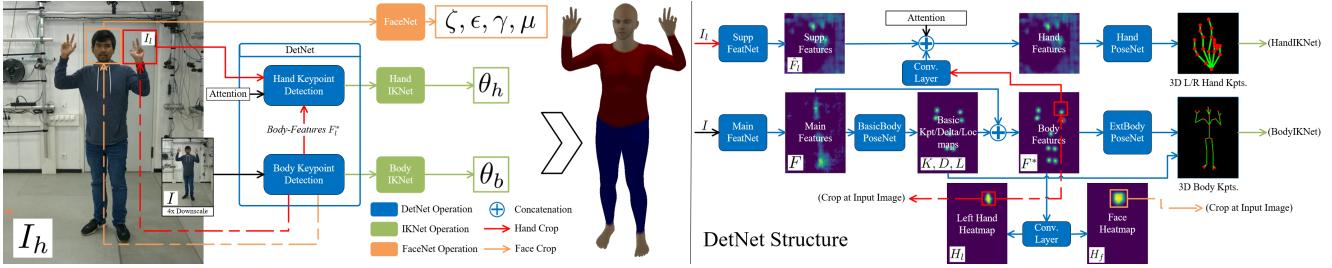


Figure 2: System overview and DetNet structure. Left: An input image I_h is first downsampled by 4x for body keypoint detection and face/hand localization. The hand area is then cropped from I_h to retrieve *supp-features*, which are concatenated with processed *body-features* for hand keypoint detection. Here, we use the attention channel to indicate the validity of *body-features*. Body and hand 3D keypoint positions are fed into *BodyIKNet* and *HandIKNet* to estimate joint angles. The face area is cropped from I_h and processed by *FaceNet*. Finally, the parameters are combined to obtain a full mesh. Right: The detailed structure of *DetNet*. Descriptions can be found in Sec. 3.2. We only illustrate one hand for simplicity.

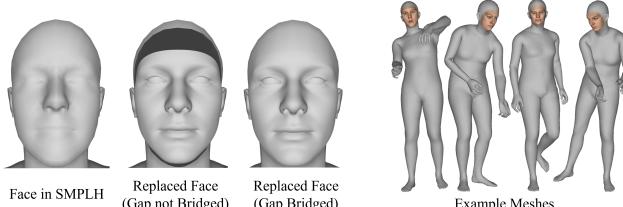


Figure 3: Our mesh model. From left to right: the original face in SMPLH; the replaced face (gap not bridged); the replaced face (gap bridged); example full body meshes.

keypoints into two subsets: *basic body keypoints* which are shared by all body datasets without annotation discrepancy, and *extended body keypoints* that are dataset-specific. We use one *BasicBody-PoseNet* to predict the *basic body keypoints* for all datasets, and use different *ExtBody-PoseNets* to estimate the *extended body keypoints* for different datasets. This separation is essential for the multi-dataset training, and avoids *BasicBody-PoseNet* to be biased to a specific dataset. The *-PoseNet* structure will be detailed in Sec. 3.2.5.

The input of *DetNet* is an image I_h of resolution 768×1024 with one person as the main subject. We bilinearly downscale it by a factor of 4 to get the low resolution image I , and feed it into the *MainFeatNet*, a ResNet [25] alike feature extractor, to obtain main features F , which are fed into *BasicBody-PoseNet* to estimate *basic body keypoints*. We then concatenate the features F with the outputs of *BasicBody-PoseNet* to get the body features F^* , which encodes high-level features and body information. Finally, we use *ExtBody-PoseNet* to predict the *extended body keypoints* from F^* . The *basic body keypoints* and *extended body keypoints* are combined to obtain the complete body keypoints.

3.2.2 Hand Localization

From the body features F^* , we use one convolutional layer to estimate left and right hand heat-maps H_l and H_r . For each hand, its heat-map H is a one-channel 2D map where the value at each pixel represents the confidence that this

pixel is occupied by the hand. We use a sliding window to locate each hand from H , determined by its width w and top-left corner location (u, v) , given by

$$\arg \min_w : \max_{u, v} \sum_{i=u, j=v}^{i < u+w, j < v+w} h_{ij} > t * \sum_{i=0, j=0}^{i < a, j < b} h_{ij} \quad (6)$$

where h_{ij} is the confidence value of H at pixel (i, j) ; a and b are the width and height of H ; and t is a manually-set threshold value. The intuition behind is to take the bounding box of minimal size that sufficiently contains the hand. This heat-map based approach is consistent with the convolutional structure and the information of body embedded in F^* is naturally leveraged in the estimation of H .

3.2.3 Hand Keypoint Detection with Attention-based Feature Composition

After hand localization, for the left and right hand, we crop F^* at the area of the hands to get the corresponding features F_l^* and F_r^* , referred to as *body-features*. They represent high-level global features. Similarly, we crop the high resolution input image I_h to get the left and right hand images I_l and I_r , which are processed by *SuppFeatNet* to obtain supplementary features \hat{F}_l and \hat{F}_r , referred to as *supp-features*. They represent high-frequency local features. For each hand, its corresponding *body-features* are bilinearly resized and processed by one convolutional layer and then concatenated with its *supp-features*. The combined features are fed into *Hand-PoseNet* to estimate hand keypoints. This feature composition exploits the inter-part correlations between body and hands, and saves the computation of high-level features of the hand area by streaming directly from the body branch. For time efficiency, *SuppFeatNet* is designed to be a shallow network with only 8 ResNet blocks. We use one *SuppFeatNet* that handles I_l and horizontally flipped I_r at the same time. The extracted features of I_r are then flipped back. On the other hand, we use two separate *Hand-PoseNets* for the two hands, as different hands focus on different channels of F^* .

To leverage hand-only datasets for training, we further introduce an *attention* mechanism that guides the hand branch to ignore *body-features* when the body is not presented in the image. Specifically, we additionally feed a one-channel binary-valued map into *Hand-PoseNet* to indicate whether the *body-features* are valid. When the body is presented in the training sample, we set it to 1; otherwise, it is set to 0. At inference, it is always set to 1.

3.2.4 Face Localization

DetNet localizes the face in the input image using a face heat-map H_f similarly as Eq. 6. The face is cropped from the input image and later used to regress the face parameters by the separately trained *FaceNet* module introduced in Sec. 3.4. Different to the hands, *FaceNet* only requires the face image and does not take F^* as input. This is based on our observation that the image input is sufficient for our fast *FaceNet* to capture the face with high quality.

3.2.5 Other Details

PoseNet Module. The *BasicBody-PoseNet*, the *ExtBody-PoseNet*, and the *Hand-PoseNet* share the same atomic network structure which comprises 6 convolutional layers to regress keypoint-maps K (for 2D keypoint positions), delta-maps D (for 3D bone directions), and location-maps L (for 3D keypoint locations) from input features. At inference, the coordinate of keypoint i is retrieved from the location-map L_i at the position of the maximum of the keypoint-map K_i . The delta-map D_i is for involving intermediate supervision. Please refer to the supplementary document and [40] for more details. The atomic loss function of this module is formulated as follows:

$$\mathcal{L}_p = w_k \mathcal{L}_{\text{kmap}} + w_d \mathcal{L}_{\text{dmap}} + w_l \mathcal{L}_{\text{lmap}} \quad (7)$$

where

$$\mathcal{L}_{\text{kmap}} = \|K^{\text{GT}} - K\|_F^2 \quad (8)$$

$$\mathcal{L}_{\text{dmap}} = \|K^{\text{GT}} \odot (D^{\text{GT}} - D)\|_F^2 \quad (9)$$

$$\mathcal{L}_{\text{lmap}} = \|K^{\text{GT}} \odot (L^{\text{GT}} - L)\|_F^2. \quad (10)$$

K , D and L are keypoint-maps, delta-maps, and location-maps, respectively. Superscript \cdot^{GT} denotes the ground truth, $\|\cdot\|_F$ is the Frobenius norm, and \odot is the element-wise product. K^{GT} is obtained by placing Gaussian kernels centered at the 2D keypoint locations. D^{GT} and L^{GT} are constructed by tiling ground truth 3D keypoint coordinates and unit bone direction vectors to the size of K^{GT} . w_k , w_d and w_l are hyperparameters to balance the terms. For the training data without 3D labels, we set w_d and w_l to 0.

Full Loss. The full loss function of the *DetNet* is defined as

$$\lambda_b \mathcal{L}_p^b + \lambda_h (\mathcal{L}_p^{lh} + \mathcal{L}_p^{rh} + \mathcal{L}_h) + \lambda_f \mathcal{L}_f. \quad (11)$$

\mathcal{L}_p^b , \mathcal{L}_p^{lh} , and \mathcal{L}_p^{rh} are the keypoint detection losses for body, left hand and right hand, respectively.

$$\mathcal{L}_h = \|H_l^{\text{GT}} - H_l\|^2 + \|H_r^{\text{GT}} - H_r\|^2 \quad (12)$$

supervises hand heat-maps for hand localization. Similarly,

$$\mathcal{L}_f = \|H_f^{\text{GT}} - H_f\|^2 \quad (13)$$

supervises the face heat-map. H_f^{GT} , H_l^{GT} , and H_r^{GT} are constructed by taking the maximum along the channel axis of the keypoint-maps to obtain a one-channel confidence map. λ_b , λ_h , and λ_f are hyperparameters which are set to 0 when the corresponding parts are not in the training sample.

Global Translation. All monocular approaches suffer from depth-scale ambiguity. In *DetNet*, the estimated keypoint positions are relative to the root keypoint. However, when the camera intrinsics matrix C and the length of any bone l_{cp} are known, the global translation can be determined based on

$$l_{cp} = \|C^{-1} z_p \begin{bmatrix} u_p \\ v_p \\ 1 \end{bmatrix} - C^{-1} (z_p + d_c - d_p) \begin{bmatrix} u_w \\ v_w \\ 1 \end{bmatrix}\|_2. \quad (14)$$

Here, the subscript \cdot_c and \cdot_p denote the child and parent keypoint of bone l_{cp} ; u and v are 2D keypoint positions; d refers to the root-relative depth; and z_p is the absolute depth of keypoint p relative to the camera. In Eq. 14, z_p is the only unknown variable that can be solved in closed form. When z_p is known, the global translation can be computed with the camera projection formula.

3.3 Inverse Kinematics Network: IKNet

Sparse 3D keypoint positions are not sufficient to drive CG character models. To animate mesh models and obtain dense surface, joint angles need to be estimated from sparse keypoints. This task is known as inverse kinematics (IK). Typically, the IK task is tackled with iterative optimization methods [6, 21, 68, 69, 22, 63], which are sensitive to initialization, take longer time, and need hand-crafted priors. Instead, we use a fully connected neural network module, referred to as *IKNet*, to regress joint angles from keypoint coordinates, similar to [78]. Trained with additional MoCap data, *IKNet* learns a pose prior implicitly from the data, and as a result further decreases keypoint position errors. Due to the end-to-end architecture, *IKNet* achieves superior run-time performance, which is crucial for being real-time.

In particular, *IKNet* is a fully connected network that takes in keypoint coordinates and outputs joint rotations θ_b and θ_h for body and hands. The main difference between our approach and [78] is that we use relative 6D rotation [77] as the output formulation, and our network additionally estimates the shape parameters β and a scale factor α . Since there is little MoCap data that contains body and

hand joint rotations simultaneously, and synthesizing such data is not guaranteed to be anatomically correct, we train *BodyIKNet* and *HandIKNet* to estimate θ_b and θ_h separately, instead of training a single network that regresses all joint angles. The loss terms are defined as:

$$\lambda_\alpha L_\alpha + \lambda_\beta L_\beta + \lambda_\theta L_\theta + \lambda_\chi L_\chi + \lambda_{\bar{\chi}} L_{\bar{\chi}}. \quad (15)$$

Here, L_α , L_β , L_θ , L_χ , and $L_{\bar{\chi}}$ are L2 losses for the scale factor α , shape parameters β , joint rotations θ , keypoint coordinates after posing χ , and keypoint coordinates at the reference pose $\bar{\chi}$. $\lambda.$ are the weights for different terms.

3.4. Face Parameters Estimation: FaceNet

We adopt a convolutional module, named *FaceNet*, to estimate shape, expression, albedo and illumination parameters of a statistical 3DMM face model [5] from a face-centered image. The face image is obtained by cropping the original high-resolution image according to the face heatmap estimated by *DetNet*. Compared with previous full body capture works [68, 46, 30, 9] that only estimate facial expression, our regression of shape, albedo and illumination gives more personalized and realistic results. *FaceNet* is originally proposed and pre-trained by Tewari et al. [61]. As the original model in [61] is sensitive to the size and location of the face in the image, we finetune it with the face crops produced by the *DetNet* for better generalization.

4. Experiments

4.1. Datasets and Evaluation Metrics

The following datasets are used to train *DetNet*: 1) body-only datasets: HUMBI [70], MPII3D [39], HM36M [26], SPIN [33], MPII2D [1], and COCO [35]; 2) hand-only datasets: FreiHand [80], STB [73], and CMU-Hand [56]; 3) body with hands dataset: MTC [30]. Here, MPII2D, COCO, and CMU-Hand only have 2D labels, but they are helpful for generalization since they are in-the-wild. Please refer to the supplementary document for more details on these datasets. We utilize AMASS [37], HUMBI and SPIN to train *BodyIKNet*, and use the MoCap data from MANO [49] to train *HandIKNet* following the method of [78]. The training data for *HandIKNet* and *BodyIKNet* are augmented as in [78]. *FaceNet* is pre-trained on the Vox-Celeb2 [10] dataset following [61], and fine-tuned with face images from MTC.

We evaluate body predictions on MTC, HM36M, MPII3D, and HUMBI, using the same protocol as in [68] (MTC, HM36M) and [40] (MPII3D). On HUMBI, we select 15 keypoints for evaluation to be consistent with other datasets, and ignore the keypoints outside the image. For hand evaluation we use MTC and FreiHand. Since not all the test images in MTC have both hands annotated, we only

evaluate on the samples where both hands are labeled, referred to as MTC-Hand. We use Mean Per Joint Position Error (MPJPE) in millimeter (mm) as the metric for body and hand pose estimation, and follow the convention of previous works to report results without (default) and with (indicated by \dagger and ‘PA’) rigid alignment by performing Procrustes analysis. As [9] outputs the SMPL mesh, we use a keypoint regressor to obtain HM36M-style keypoint predictions, similar to [33, 31]. We evaluate *FaceNet* on the face images cropped from MTC test set by using 2D landmark error and per channel photometric error as the metric. We use PnP-RANSAC [16] and PA alignment to estimate camera pose for projection and error computation of the face.

4.2. Qualitative Results

We present qualitative results in Fig. 4 and compare with the state-of-the-art approach of Choutas et al. [9]. Despite much faster inference speed, our model gives results with equal visual quality. In the first row we show that our model captures detailed hand poses while [9] gives over-smooth estimation. This is because of our utilization of high-frequency local features extracted from the high-resolution hand image. In the second row, we demonstrate that our hand pose is consistent with the wrist and arm, while the result of [9] is anatomically incorrect. This is due to our utilization of body information for hand pose estimation. We demonstrate in the third row that with variations in facial shape and color, our approach provides highly personalized capture results, while [9] lacks identity information. In Fig. 5 we compare the face capture results of coarse and tight face crops. The result on the loosely cropped image already captures the subject very well (left), and a tighter bounding box obtained from a third party face detector [32] based on the coarse crop further improves the quality (right). Unless specified, the presented results in the paper are all based on tight face crops. As our approach does not estimate camera pose, for overlay visualization, we adopt PnP-RANSAC [16] and PA alignment to align our 3D and 2D predictions. The transformations are rigid and no information of ground truth is used. Please refer to the supplemental material for more results.

4.3. Quantitative Results

Runtime. Runtime performance is crucial for a variety of applications, thus real-time capability is one of our main goals. In Tab. 1, we report the runtime of each subtask in milliseconds (ms) on a commodity PC with an Intel Core i9-10920X CPU and an Nvidia 2080Ti GPU. We use -B and -H to indicate body and hand sub-tasks. Due to the efficient inter-part feature composition, it takes only 10.3ms to estimate keypoint positions of two hands, which is two times faster than the lightweight method of [78]. The end-to-end *IKNet* takes 2.68ms in total, which is nearly impossible for

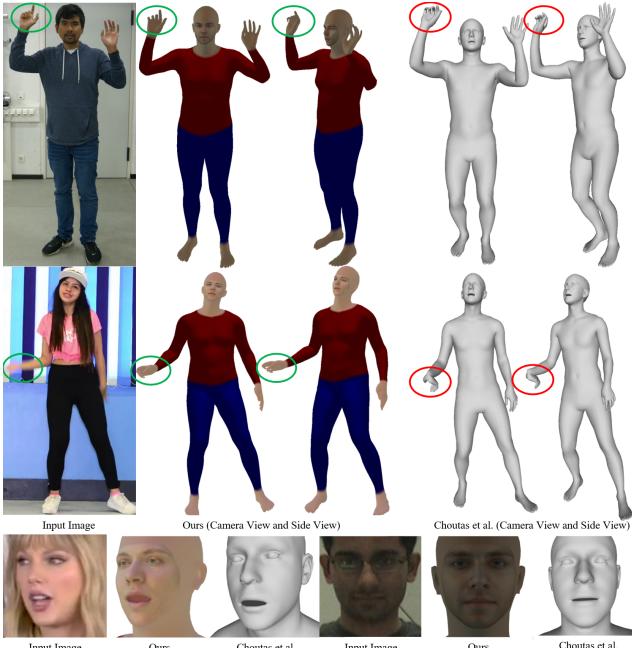


Figure 4: Qualitative results. From top to bottom: 1) our method captures subtle gestures while [9] is over-smooth; 2) our hand pose is consistent with the wrist and arm while [9] is anatomically incorrect; 3) our faces are more personalized and realistic due to the variation in identity-dependent facial geometry and albedo.

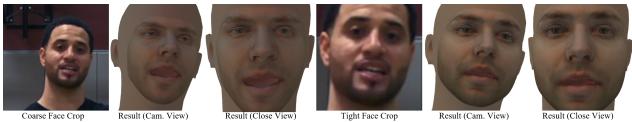


Figure 5: Comparison on face crop. A coarse face crop is already sufficient for face capture, while a tighter one further improves quality.

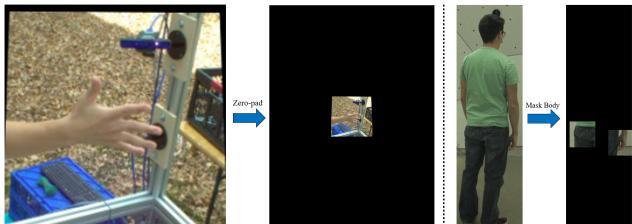


Figure 6: Samples from test data. Left: we zero-pad the hand-only image from FreiHand to evaluate our model, which is disadvantageous for us. Right: we mask the body and only keep the hand regions visible to construct the MTC-Hand-Mask test set.

traditional iterative optimization-based IK solvers. The optional face detector [32] takes 7ms, without breaking the real-time limitation (25.5fps).

Body Pose Estimation. In Tab. 2, we report quantitative evaluation for body keypoint detection of *DetNet*, and compare with other state-of-the-art approaches. Despite *DetNet* is extremely fast, it is still comparable with the top models

| Module | DetNet-B | DetNet-H | IKNet-B | IKNet-H | FaceNet | Total |
|---------|-------------|---------------|-------------|------------|---------------|-------|
| Runtime | 16.9 | 10.3 | 1.51 | 1.17 | 1.92 | 32.1 |
| Method | Ours | Kanazawa [31] | Choutas [9] | Xiang [68] | Pavlakos [46] | |
| Runtime | 32.1 | 60 | 160 | 20000 | ~50000 | |
| FPS | 31.1 | 16.7 | 6.25 | 0.05 | ~0.02 | |

Table 1: Runtime analysis in milliseconds and frames per second (FPS). Top: runtime of each subtask in our method. Bottom: comparison with previous works.

| Method | MPJPE (mm) | | | |
|------------------------|-------------------------|-------------------------|-------------------------|-------------------------|
| | HM36M | MPII3D | MTC | HUMBI |
| Xiang et al. [68] | 58.3 | - | 63.0 | - |
| Kolotouros et al. [33] | 41.1[‡] | 105.2 | - | 101.7 ^{‡§} |
| Choutas et al. [9] | 54.3 [‡] | - | - | 67.2 ^{‡§} |
| Kanazawa et al. [31] | 56.8 [‡] | 124.2 | - | 84.2 ^{‡§} |
| DetNet | 64.8 | 116.4 | 66.8 | 43.5 |
| DetNet (PA) | 50.3 [‡] | 77.0[‡] | 61.5[‡] | 32.5[‡] |

Table 2: Body MPJPE on public datasets. Our model has competitive results across all datasets while being much faster. [‡] means the model is not trained on the train split. [§] means the model is not trained on the train split.

| Metric | DetNet | DetNet+IKNet (IK-β) | DetNet+IKNet (GT- β) |
|------------|-------------------|---|-----------------------------|
| MPJPE | 43.5 | 43.3 | 39.9 |
| MPJPE (PA) | 32.5 [‡] | 31.6 [‡] | 31.2[‡] |

Table 3: Body MPJPE on HUMBI. We demonstrate that incorporating *BodyIKNet* further lowers error. The small gap between IK- β and GT- β indicates the high accuracy of body shape estimation.

in terms of accuracy. We also evaluate previous works on HUMBI although they were not trained on the train split. Notably, their accuracy significantly drops as their generalization across datasets is limited. In contrast, our approach performs similarly well across all datasets due to the multi-dataset training, indicating a better generalization ability. In Tab. 3, we compare the results after *BodyIKNet* on HUMBI with different sources of shape parameters: IK- β uses the shape parameters estimated by *BodyIKNet*, and GT- β uses the ground truth shape parameters. Due to the additional knowledge of the pose prior learned from MoCap data, *BodyIKNet* decreases the keypoint error. After PA alignment, the error of IK- β is very close to GT- β , indicating that the body shape estimation is also accurate.

Hand Pose Estimation. We report our results for hand pose estimation in Tab. 4. The results after IK are based on the shape parameters estimated by *HandIKNet*. On the MTC-Hand test set, our mean error is only 9.3mm. We attribute the 1.1mm increase of error after IK to the difference in keypoint definitions between our hand model (SMPLH) and the MTC hand model, as the bone length difference is 25% on average. When it comes to FreiHand, our error increases. This is because FreiHand is a hand-only dataset, while in our method hand pose deeply relies on body information. Since we do not have a hand-specific module, to evaluate on FreiHand, we have to zero-pad the hand image to the full size and feed it into the model (Fig. 6) as if body is presented. Despite this non-ideal setup, after IK, our error

| Method | MPJPE (mm) | | |
|----------------------------|-------------------------|-------------------------|-------------------------|
| | MTC-Hand (left) | MTC-Hand (right) | FreiHand |
| Choutas et al. [9] | 13.0 [‡] | 12.2 [‡] | 12.2[‡] |
| Zhou et al. [78] | 16.1 [‡] | 15.6 [‡] | 21.8 [‡] |
| DetNet | 15.1 | 13.8 | - |
| DetNet (PA) | 8.50[‡] | 7.90[‡] | 24.2 [‡] |
| DetNet + IKNet (PA) | 9.42 [‡] | 9.10 [‡] | 15.7 [‡] |

Table 4: Hand MPJPE on public datasets. Our model has the lowest error on MTC-Hand where the body information is available, and is comparable on FreiHand even the body is absent. [‡] means the model is not trained on the train split.

| Metric | Tewari et al. [61] | FaceNet | FaceNet-T |
|------------------|--------------------|---------|---------------|
| Landmark Err. | 4.70 | 3.43 | 3.37 |
| Photometric Err. | 0.0661 | 0.0447 | 0.0444 |

Table 5: Landmark error in pixel and photometric error per channel on MTC-Face. *FaceNet* performs better than [61] on these challenging samples, and a tighter bounding box further improves accuracy.

is still comparable to [9], and outperforms [78] which is not trained on FreiHand. Note that the previous methods in Tab. 4 are not trained on the train split of MTC and cannot compare with us directly on MTC-Hand.

Face Capture. In Tab. 5, we evaluate *FaceNet* on the face crops from the MTC test set (MTC-Face). Compared with typical datasets, the faces in MTC-Face are more blurry and challenging. Our *FaceNet* gives better results than [61] on such in-the-wild samples, and a tighter face bounding box (denoted by postfix “T”) further lowers error. Please refer to the supplementary document for more evaluation on face.

4.4. Ablation Study

Feature Composition. The inter-part feature composition from body to hands is critical to reduce runtime and improve hand pose accuracy. To examine this design, we train the following models for comparison: 1) *DetNet-S*(supplementary) where the hand branch estimates hand pose only from *supp-features* \hat{F} and does not take any information from body except hand localization; 2) *DetNet-B*(ody) where the hand branch estimates hand pose only from *body-features* F^* and does not see the high-resolution input image. To further examine the importance of body information for hand keypoint detection, we additionally construct a test set derived from MTC-Hand, called MTC-Hand-Mask, where the body area is masked and only the hands are visible (Fig. 6). The results are reported in Tab. 6. On MTC-Hand, because of the utilization of body information, the error of *DetNet* is lower than *DetNet-S* by 28%. When it comes to FreiHand and MTC-Hand-Mask, the gap between *DetNet* and *DetNet-S* shrinks to 4% and -5%. This is due to the missing body information in these two test sets, which indicates that the *body-features* indeed contribute to the hand keypoint detection. *DetNet-B* always performs worse than *DetNet*. This is because *body-features* are extracted from the low-resolution image where the hands are

| Method | MPJPE (mm) | | |
|-------------------------|-------------|---------------|-------------------------|
| | MTC-Hand | MTC-Hand-Mask | FreiHand |
| DetNet-S(supplementary) | 18.4 | 31.7 | 23.1[‡] |
| DetNet-B(ody) | 17.2 | 37.5 | 26.8 [‡] |
| DetNet | 14.4 | 30.6 | 24.2 [‡] |

Table 6: Ablation study on *body-features* and *supp-features*. The comparison between the three versions demonstrates the help of F^* and \hat{F} in the hand pose estimation task.

| Method | MPJPE (mm) | | | | |
|---------------------|-------------------------|-------------------------|-------------|-------------|-------------|
| | HM36M | MPII3D | MTC | HUMBI | MTC-Hand |
| DetNet-U(niform) | 57.9 [‡] | 99.9 [‡] | 64.6 | 59.1 | 14.7 |
| DetNet-O(verfitted) | 272.2 [‡] | 297.9 [‡] | 67.7 | 289.4 | 13.8 |
| DetNet-I(ndoor) | 61.7 [‡] | 95.7 [‡] | 64.8 | 63.1 | 15.1 |
| DetNet | 57.5[‡] | 90.1[‡] | 66.8 | 52.5 | 14.4 |

Table 7: Ablation study on training data. The gap between *DetNet-U* and *DetNet* shows the help of the attention mechanism. *DetNet-O* and *DetNet-I* only perform well on a few datasets, while *DetNet* has the best cross-dataset accuracy.

too blurry and cover only a few pixels. This comparison indicates the importance of *supp-features*.

Data Modalities. The advantage of using MoCap data is examined in Tab. 3 where *IKNet* lowers the error. To evaluate the attention mechanism and multiple image datasets, we train the following models: 1) *DetNet-U*(niform) which is trained without the attention mechanism, i.e. we treat hand-only data as if body is presented by always setting the attention channel to 1; 2) *DetNet-O*(verfitted) which is trained on the only dataset where body and hands are annotated simultaneously, namely MTC; 3) *DetNet-I*(ndoor) that only uses the training data with 3D annotations (usually indoor) without any 2D-labeled data (usually in-the-wild). To account for different keypoint definitions, we only evaluate *basic body keypoints*, except for MTC where all the models are trained on. As shown in Tab. 7, *DetNet-U* generally performs worse than *DetNet*, indicating that the attention mechanism helps during training. *DetNet-O* has poor cross-dataset generalization and only performs well on MTC-Hand. This illustrates the importance of the multi-dataset training strategy, which is enabled by our 2-stage keypoint detection structure. Finally, the inferior of *DetNet-I* to *DetNet* demonstrates the help of in-the-wild images, although they only have 2D annotations. Please refer to the supplementary video for more evaluation on the training data.

5. Conclusion

We present the first real-time approach to capture body, hands, and face from an RGB image. The accuracy and time efficiency comes from our network design that exploits inter-part relationship between body and hands. By training the network as separate modules, we leverage multiple data sources and achieve superior generalization. Further, our approach captures personalized face with both expression and identity-dependent shape and albedo. Future directions can involve temporal information for smoother results.

References

- [1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3686–3693, 2014.
- [2] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*, pages 408–416. 2005.
- [3] Bruno Artacho and Andreas Savakis. Unipose: Unified human pose estimation in single images and videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [4] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Weakly-supervised domain adaptation via gan and mesh model for estimating 3d hand poses interacting objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6121–6131, 2020.
- [5] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999.
- [6] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European Conference on Computer Vision*, pages 561–578. Springer, 2016.
- [7] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10843–10852, 2019.
- [8] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*, 2018.
- [9] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Monocular expressive body regression through body-driven attention. *arXiv preprint arXiv:2008.09062*, 2020.
- [10] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018.
- [11] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018.
- [12] Matthias Dantone, Juergen Gall, Christian Leistner, and Luc Van Gool. Human pose estimation using body parts dependent joint regressors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3041–3048, 2013.
- [13] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David J Crandall. Hope-net: A graph-based model for hand-object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6608–6617, 2020.
- [14] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (TOG)*, 39(5):1–38, 2020.
- [15] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2334–2343, 2017.
- [16] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.
- [17] Valentin Gabeur, Jean-Sébastien Franco, Xavier Martin, Cordelia Schmid, and Gregory Rogez. Moulding humans: Non-parametric 3d human shape estimation from single images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2232–2241, 2019.
- [18] Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. Reconstruction of personalized 3d face rigs from monocular video. *ACM Transactions on Graphics (TOG)*, 35(3):1–15, 2016.
- [19] Liuhan Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 10833–10842, 2019.
- [20] Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. Multiview face capture using polarized spherical gradient illumination. In *Proceedings of the 2011 SIGGRAPH Asia Conference*, pages 1–10, 2011.
- [21] Peng Guan, Alexander Weiss, Alexandru O Balan, and Michael J Black. Estimating human shape and pose from a single image. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1381–1388. IEEE, 2009.
- [22] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. Livecap: Real-time human performance capture from monocular video. *ACM Transactions on Graphics (TOG)*, 38(2):1–17, 2019.
- [23] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5052–5063, 2020.
- [24] Ihsanul Habibie, Weipeng Xu, Dushyant Mehta, Gerard Pons-Moll, and Christian Theobalt. In the wild human pose estimation using explicit 2d features and intermediate 3d representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10905–10914, 2019.
- [25] 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.
- [26] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environ-

- ments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013.
- [27] Aaron S Jackson, Chris Manafas, and Georgios Tzimiropoulos. 3d human body reconstruction from a single image via volumetric regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.
- [28] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. *arXiv preprint arXiv:2007.11858*, 2020.
- [29] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart Nabbe, Iain Matthews, et al. Panoptic studio: A massively multiview system for social interaction capture. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):190–204, 2017.
- [30] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8320–8329, 2018.
- [31] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018.
- [32] Davis E King. Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, 10:1755–1758, 2009.
- [33] Nikos Kolotouros, Georgios Pavlakos, Michael Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2252–2261, 2019.
- [34] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4501–4510, 2019.
- [35] 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.
- [36] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015.
- [37] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5442–5451, 2019.
- [38] Gines Hidalgo Martinez, Yaadhav Raaj, Haroon Idrees, Donglai Xiang, Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Single-network whole-body pose estimation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6981–6990, 2019.
- [39] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 international conference on 3D vision (3DV)*, pages 506–516. IEEE, 2017.
- [40] 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. *international conference on computer graphics and interactive techniques*, 36(4):1–14, 2017.
- [41] Gyeongsik Moon, Takaaki Shiratori, and Kyoung Mu Lee. Deephandmesh: A weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling. *arXiv preprint arXiv:2008.08213*, 2020.
- [42] 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 *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, volume 271, pages 49–59, 2018.
- [43] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1284–1293, 2017.
- [44] Evonne Ng, Hanbyul Joo, Shiry Ginosar, and Trevor Darrell. Body2hands: Learning to infer 3d hands from conversational gesture body dynamics. *arXiv preprint arXiv:2007.12287*, 2020.
- [45] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *2018 International Conference on 3D Vision (3DV)*, pages 484–494, 2018.
- [46] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019.
- [47] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning detailed face reconstruction from a single image. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1259–1268, 2017.
- [48] Sami Romdhani and Thomas Vetter. Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, volume 2, pages 986–993. IEEE, 2005.
- [49] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6):245, 2017.
- [50] Joseph Roth, Yiying Tong, and Xiaoming Liu. Adaptive 3d face reconstruction from unconstrained photo collections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4197–4206, 2016.
- [51] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morigi, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned

- implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2304–2314, 2019.
- [52] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020.
- [53] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D. Castillo, and David W. Jacobs. Sfsnet: Learning shape, reflectance and illuminance of faces ‘in the wild’. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6296–6305, 2018.
- [54] Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. Physcap: Physically plausible monocular 3d motion capture in real time. *arXiv preprint arXiv:2008.08880*, 2020.
- [55] Leonid Sigal and Michael J Black. Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’06)*, volume 2, pages 2041–2048. IEEE, 2006.
- [56] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1145–1153, 2017.
- [57] David Smith, Matthew Loper, Xiaochen Hu, Paris Mavroidis, and Javier Romero. Facsimile: Fast and accurate scans from an image in less than a second. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5330–5339, 2019.
- [58] Bugra Tekin, Isinsu Katircioglu, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. Structured prediction of 3d human pose with deep neural networks. *arXiv preprint arXiv:1605.05180*, 2016.
- [59] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zöllhofer, and Christian Theobalt. Fml: Face model learning from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10812–10822, 2019.
- [60] Ayush Tewari, Michael Zollhofer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Perez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2549–2559, 2018.
- [61] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 1274–1283, 2017.
- [62] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016.
- [63] Anastasia Tkach, Mark Pauly, and Andrea Tagliasacchi. Sphere-meshes for real-time hand modeling and tracking. *ACM Transactions on Graphics (ToG)*, 35(6):1–11, 2016.
- [64] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5163–5172, 2017.
- [65] Yangang Wang, Cong Peng, and Yebin Liu. Mask-pose cascaded cnn for 2d hand pose estimation from single color image. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(11):3258–3268, 2018.
- [66] Zhibo Wang, Jingwang Ling, Chengzeng Feng, Ming Lu, and Feng Xu. Emotion-preserving blendshape update with real-time face tracking. *IEEE Annals of the History of Computing*, (01):1–1, 2020.
- [67] Philippe Weinzaepfel, Romain Brégier, Hadrien Combaluzier, Vincent Leroy, and Grégory Rogez. Dope: Distillation of part experts for whole-body 3d pose estimation in the wild. In *ECCV*, 2020.
- [68] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10965–10974, 2019.
- [69] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. Monopercap: Human performance capture from monocular video. *ACM Transactions on Graphics (ToG)*, 37(2):1–15, 2018.
- [70] Zhixuan Yu, Jae Shin Yoon, In Kyu Lee, Prashanth Venkatesh, Jaesik Park, Jihun Yu, and Hyun Soo Park. Humbi: A large multiview dataset of human body expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2990–3000, 2020.
- [71] Shanxin Yuan, Guillermo Garcia-Hernando, Björn Stenger, Gyeongsik Moon, Ju Yong Chang, Kyoung Mu Lee, Pavlo Molchanov, Jan Kautz, Sina Honari, Liuhao Ge, et al. Depth-based 3d hand pose estimation: From current achievements to future goals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2636–2645, 2018.
- [72] Hao Zhang, Zi-Hao Bo, Jun-Hai Yong, and Feng Xu. Interactionfusion: real-time reconstruction of hand poses and deformable objects in hand-object interactions. *ACM Transactions on Graphics (TOG)*, 38(4):1–11, 2019.
- [73] Jiawei Zhang, Jianbo Jiao, Mingliang Chen, Liangqiong Qu, Xiaobin Xu, and Qingxiong Yang. A hand pose tracking benchmark from stereo matching. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 982–986, 2017.
- [74] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2354–2364, 2019.

- [75] Yuxiang Zhang, Liang An, Tao Yu, Xiu Li, Kun Li, and Yebin Liu. 4d association graph for realtime multi-person motion capture using multiple video cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1324–1333, 2020.
- [76] Zerong Zheng, Tao Yu, Yixuan Wei, Qionghai Dai, and Yebin Liu. DeepHuman: 3d human reconstruction from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7739–7749, 2019.
- [77] Yi Zhou, Connnelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019.
- [78] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5346–5355, 2020.
- [79] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE international conference on computer vision*, pages 4903–4911, 2017.
- [80] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max J. Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 813–822, 2019.
- [81] Michael Zollhöfer, Justus Thies, Pablo Garrido, Derek Bradley, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. State of the art on monocular 3d face reconstruction, tracking, and applications. In *Computer Graphics Forum*, volume 37, pages 523–550. Wiley Online Library, 2018.