HiFECap: Monocular High-Fidelity and Expressive Capture of Human Performances

Yue Jiang
yue.jiang@aalto.fi

Marc Habermann
Vladislav Golyanik
Christian Theobalt
{mhaberma,golyanik,theobalt}@mpi-inf.mpg.de

Max Planck Institute for Informatics
Saarland Informatics Campus

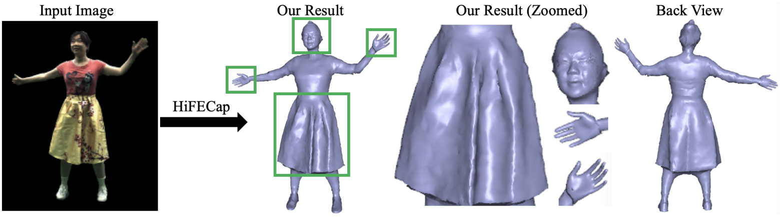

Figure 1: We present a new monocular 3D human performance capture approach, HiFECap, which for the first time jointly captures the body pose, hand gestures, facial expressions, *and* high-frequency non-rigid deformations in 3D solely using RGB images as input. The deformations recovered by our method are a clear step towards higher-fidelity cloth capture.

Abstract

Monocular 3D human performance capture is indispensable for many applications in computer graphics and vision for enabling immersive experiences. However, detailed capture of humans requires tracking of multiple aspects, including the skeletal pose, the dynamic surface, which includes clothing, hand gestures as well as facial expressions. No existing monocular method allows joint tracking of all these components. To this end, we propose HiFECap, a new neural human performance capture approach, which simultaneously captures human pose, clothing, facial expression, and hands just from a single RGB video. We demonstrate that our proposed network architecture, the carefully designed training strategy, and the tight integration of parametric face and hand models to a template mesh enable the capture of all these individual aspects. Importantly, our method also captures high-frequency details, such as deforming wrinkles on the clothes, better than the previous works. Furthermore, we show that HiFECap outperforms the state-of-the-art human performance capture approaches qualitatively and quantitatively while for the first time capturing all aspects of the human.

1 Introduction

The goal of 3D human performance capture is the space-time coherent tracking of the entire human surface from different sensor types; this is a long-standing and challenging computer vision problem. Such densely tracked characters can be used in film, game, and mixed reality applications to create immersive and photo-real virtual doubles of real humans.

Previous multi-view-based approaches [11, 12, 13, 14, 19, 22, 48, 54, 58, 59, 79, 84] can capture high-quality surface details. However, they rely on impractical and expensive multi-camera capture setups. To commoditize performance capture, ideally, just a single RGB camera should be necessary while still allowing users to track both the body pose and non-rigid deformations of skin and clothing. Prior monocular approaches were able to recover the pose and shape of a naked human body model [38, 39, 56], hands [10, 17, 23, 52, 81, 92, 93, 96], facial expression [41, 71, 72, 73, 76], or all of them [36, 57, 86, 97]; recovering cloth deformations remains out of their reach. Some previous work on monocular 3D human and clothes reconstruction uses volumetric [21, 94] or continuous implicit representations [64]. However, these approaches do not track space-time coherent surfaces and lack surface correspondences over time. On the other hand, template-based monocular methods [24, 25, 46, 89] can track low-frequency surface details coherently over time, but cannot capture facial expressions and hand gestures. Joint capture of all aspects remains poorly studied.

To address these limitations, we present *HiFECap*, a novel monocular learning-based 3D human performance capture approach that jointly captures the skeletal pose, dense surface deformations, hand gestures, and facial identity and expressions; see Fig. 1 for an overview. First, convolutional neural networks predict the skeletal pose and the coarse surface deformations from the segmented monocular image of the actor. High-frequency surface details are recovered by a deformation network as dense vertex displacements. These intermediate outputs are then combined in a differentiable character representation, which can be supervised with multi-view images and 3D point clouds during training. We further replace the hand and face regions of the original template with parametric hand and face models using our proposed registration strategy, and drive them by predicting the parameters from images.

In summary, our **technical contributions** are: 1) HiFECap, *i.e.,* the first monocular 3D human performance capture approach enabling joint tracking of body pose, the non-rigidly deforming surface, hand gestures, and facial expressions. 2) A visibility- and rigidity-aware vertex displacement network to enable the capture of high-frequency geometric details of the dynamic human surface. 3) A multi-stage training process for surface recovery and a face and hand model integration. Our experiments show that HiFECap applies to different clothing types and outperforms the existing state of the art in terms of recovered details.

2 Related Work

Multi-view Performance Capture. Many approaches require multi-view imagery [18, 50, 67, 80, 82]. Prior works reconstruct surface deformations based on person-specific template meshes [14, 16, 19] or a volumetric representation [2, 31]. For high-quality reconstructions, some methods rely on segmented and high-resolution human body scans [11, 13, 48, 83] or articulated skeletons to separate piece-wise rigid and non-rigid deformations [22, 48, 79, 84]. Parametric models offer another possibility to 3D human motion capture [4, 27, 30, 37, 42, 49, 55, 68]. Approaches relying on them often ignore clothing by treating it as noise [6], or estimate only naked body shape [5, 32, 90, 91]. Some techniques track facial expressions [36] and hands [36, 63] in addition to the proxy human body shape.

To capture the human with clothing, some methods deform a 3D model to fit a scan [91] or multi-view images [61], use separate meshes for body shape and clothing [59] or deploy multi-view CNNs [33]. However, all these approaches require a multi-view setup *at inference time*, which makes them impractical for most users. In contrast, our method only leverages a multi-view setting *for capturing training data*. Once our high-fidelity, expressive, and personalized approach is trained, it only takes a single RGB video as input at inference time.

Monocular 3D Pose Estimation and Performance Capture. Monocular performance capture is an ill-posed problem with lots of ambiguities (e.g., along the depth dimension and due to occlusions). Leveraging 2D and 3D joint detections, many methods capture 3D human motion from monocular images by predicting 3D poses [51, 60, 62, 69, 70, 74, 95] or fit a parametric body model [3, 8, 38, 44, 45, 49, 78]. Other methods directly regress the body model parameters [38, 39, 56] or a coarse volumetric body shape [77], and can also jointly capture body pose with facial expressions and hand gestures [20, 57, 86, 97]. PIFuHD [65] and SelfRecon [34] work for standing poses and do not generalize to arbitrary poses. Moreover, PIFu[HD] [64, 65] reconstructs per-frame geometry, while our work aims at tracking a space-time coherent geometry, which by nature is in correspondence over time.

Capturing the non-rigid and dynamic surface of the person's clothing from monocular videos remains challenging. MonoClothCap [87] estimates the deforming surface without the need of a person-specific template. Instead, they deform a parametric body model during capture. However, they cannot track clothing types with a topology that is significantly different from the body model, e.g. skirts and dresses.

Most closely related to our work are template-based monocular 3D human performance capture methods [24, 25, 46, 88, 89]. MonoPerfCap [89] tracks an actor observed in a monocular video using a 3D actor's template. This method is based on global energy optimisation, and, hence, its runtime is high and the results appear oversmoothed in many cases. In contrast, LiveCap [24] achieves real-time performance and DeepCap [25] further improves 3D accuracy by employing a neural architecture and using multi-view supervision during training. Further, replacing the geometric surface regularization (e.g., as-rigid-as-possible regulariser) with a physics-based constraint improves the physical plausibility of the deformations [46]. All of the above-mentioned methods cannot regress facial expressions, hand gestures, and high-frequency deformations. In contrast, our HiFECap approach captures the state-specific appearance of the face and hands and high-frequency surface details—for the first time—in a single framework for expressive 3D human performance capture.

3 Method

Given a monocular video of a human in motion, the goal of our method is to regress the 3D deformation of a person-specific template mesh of the human including clothing, hand gestures, and facial expressions for each of the video frames. To this end, we first acquire multi-view training images of the human performing a diverse set of actions and define a differentiable character representation, which efficiently parameterizes the template from coarse to fine (Sec. 3.1). Then, we propose regression networks in a coarse-to-fine manner, *i.e.,* we first employ a skeletal pose prediction network and a coarse embedded deformation network, which captures the piece-wise rigid skeletal deformations and the coarse surface deformations, respectively (Sec. 3.2). For capturing finer surface details, we propose a novel hybrid image-to-graph convolutional architecture for predicting per-vertex displacements, which greatly improves the dynamic surface details (Sec. 3.3). Since supervised learning of the network models is not possible, we resume to a weakly-supervised setup and propose a carefully designed combination of loss functions all geared towards high fidelity surface capture (Sec. 3.4). Last, we replace the face and hand regions of the original template with parametric models using our proposed registration procedure. Then, a dedicated network is predicting the facial expression as well as the hand gestures (Sec. 3.5).

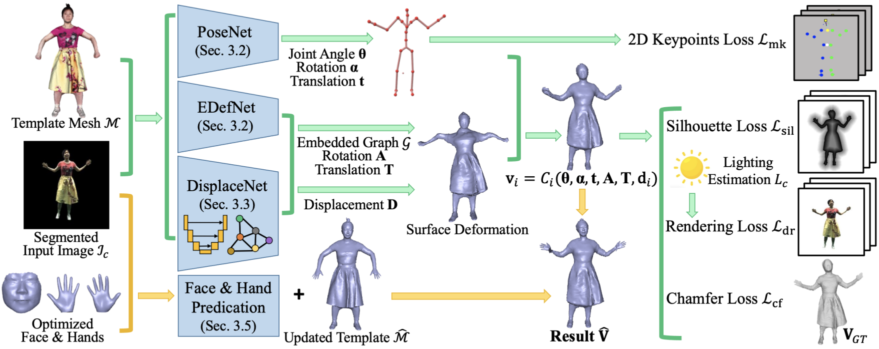
Template Mesh $\mathcal{M}$
Segmented Input Image $\mathcal{I}_c$
Optimized Face & Hands
PoseNet (Sec. 3.2)
EDefNet (Sec. 3.2)
DisplaceNet (Sec. 3.3)
Face & Hand Prediction (Sec. 3.5)
Joint Angle $\boldsymbol{\theta}$ Rotation $\boldsymbol{\alpha}$ Translation $\mathbf{t}$
Embedded Graph $\mathcal{G}$ Rotation $\mathbf{A}$ Translation $\mathbf{T}$
Displacement $\mathbf{D}$
Surface Deformation
+
Updated Template $\hat{\mathcal{M}}$
$\mathbf{v}_i = C_i(\boldsymbol{\theta}, \boldsymbol{\alpha}, \mathbf{t}, \mathbf{A}, \mathbf{T}, \mathbf{d}_i)$
Result $\hat{\mathbf{V}}$
2D Keypoints Loss $\mathcal{L}_{\mathrm{mk}}$
Silhouette Loss $\mathcal{L}_{\mathrm{sil}}$
Lighting Estimation L_c
Rendering Loss $\mathcal{L}_{\mathrm{dr}}$
Chamfer Loss $\mathcal{L}_{\mathrm{cf}}$
$\mathbf{V}_{GT}$

Figure 2: **Overview of our HiFECap approach** that takes a single segmented image as input and tracks the corresponding 3D human mesh. *PoseNet* estimates the 3D skeletal pose as joint angles and a global rotation. It is followed by coarse-to-fine deformation regression based on silhouette, rendering and Chamfer losses. *EDefNet* captures coarse skin and clothing details by predicting the deformation on the embedded graph. *DisplaceNet* refines the results with high-frequency details based on a vertex displacement field (green arrows). We then replace the corresponding template parts with parametric hand and face models. Given the input image, a dedicated network then predicts those parameters (yellow arrows).

3.1 Data Processing and Character Representation

For training, we record a multi-view video of the actor performing various motions in a studio with a green screen background. We detect 2D joint keypoints using OpenPose [15], apply color keying to extract foreground masks, and generate respective distance transform images [9] for each view and frame. We use a multi-view stereo reconstruction software Agisoft Metashape [1] to reconstruct the ground truth mesh $\mathbf{V}_{\mathrm{GT},f}$ for each frame. As input to our method, we randomly sample cropped and segmented frames $\mathcal{I}_{f,c}$ where f and c denote the frame and camera index. At the same time, other views of frame f are used for supervision. For testing, we record in-the-wild monocular videos, extract the foreground masks using Detectron2 [85], and use OpenPose for retrieving 2D keypoint detections. For simplicity, we omit the subscript f in the following.

Our method requires a person-specific textured, rigged, and skinned 3D template of the actor. Therefore, we scan the person in a multi-view stereo scanner [75] and use Metashape [1] to reconstruct the 3D mesh with around $N{\approx}5000$ vertices. We rig the scanned mesh to a kinematic skeleton being parameterized with the root rotation $\alpha \in \mathbb{R}^3$, the global translation $\mathbf{t} \in \mathbb{R}^3$, and the joint angles $\theta \in \mathbb{R}^{33}$. We also attach 3D landmarks to the skeleton (21 body joints and 6 face landmarks). We automatically compute the skinning weights in Blender [29] and leverage Dual Quaternion Skinning [40] to deform the mesh based on the skeletal pose. Similar to Habermann et al. [24], we assign a rigidity weight r_i to each vertex $\mathbf{V}_i$ to account for the different deformation properties of varying materials. We further define a downsampled version of the mesh as the underlying embedded graph $\mathcal{G}$ and model deformations from coarse to fine using an embedded deformation [68]. $\mathcal{G}$ is parameterized with $\mathbf{A} \in \mathbb{R}^{K\times 3}$ and $\mathbf{T} \in \mathbb{R}^{K\times 3}$ representing local graph rotations and translations, respectively. Here, K denotes the number of graph nodes. To capture finer high-frequency geometric details of non-rigid deformations such as garment folds, we use a 3D vertex displacement map $\mathbf{D}$, *i.e.,* we assign a displacement vector $\mathbf{d}_i \in \mathbb{R}^3$ to each mesh vertex. Similar to DDC [26], the final character representation $\mathbf{v}_i = C_i(\theta, \alpha, \mathbf{t}, \mathbf{A}, \mathbf{T}, \mathbf{d}_i)$, internally applies the embedded deformation and the vertex displacements in a canonical T-pose based on the parameters $\mathbf{A}, \mathbf{T}, \mathbf{d}_i$ and finally poses the deformed mesh based on the skeletal pose $\theta, \alpha, \mathbf{t}$.

3.2 Pose Network and Embedded Deformation Network

Similar to DeepCap [25], *PoseNet* and *EDefNet* are both Resnet50-based networks [28] that take the image $\mathcal{I}_c$ of view c as input. The network architecture of *EDefNet* is the same as *DefNet* in DeepCap [25], but we train it using additional loss terms in addition to the 2D supervision proposed in DeepCap (details in supplemental document). *PoseNet* regresses the skeleton joint angles $\theta \in \mathbb{R}^{27}$ and the camera relative root rotations $\alpha \in \mathbb{R}^3$. The global translation of the mesh template $\mathbf{t} \in \mathbb{R}^3$ is obtained by a global alignment layer [25]. We supervise *PoseNet* with a multi-view 2D keypoint loss and a joint angle regularizer as proposed in DeepCap [25]. *EDefNet* regresses the embedded deformation parameters $\mathbf{A}, \mathbf{T}$, which capture the coarse surface deformations.

3.3 Visibility- and Rigidity-aware Vertex Displacement Network

To capture high-frequency geometric details, we add a per-vertex displacement network, *DisplaceNet*, which takes the input image $\mathcal{I}_c$ and regresses the vertex displacement field $\mathbf{D} \in \mathbb{R}^{N\times 3}$ in the canonical pose ($\mathbf{d}_i$ denotes the i-th row of $\mathbf{D}$). For this task, we found that local image patches usually contain most of the relevant information about the high-frequency deformation patterns. Thus, we introduce a novel architecture, which maps local image features onto a graph convolutional architecture to improve accuracy and robustness.

Image Feature Map Encoder. First, we use a U-Net-based image encoder *DUNet*, which extracts relevant information about the surface deformation, *e.g.*, wrinkles from the input image. More precisely, it takes the input frame $\mathcal{I}_c \in \mathbb{R}^{256\times 256\times 3}$ and computes a latent feature map $f_{\text{DUNet},c}(\mathcal{I}_c) = \mathcal{F}_c \in \mathbb{R}^{256\times 256\times 32}$ with the same spatial resolution as the input frame.

Visibility-aware Vertex Feature Map. Next, those features in image space are mapped onto the posed and coarsely deformed mesh: We propose a function $P_c(\mathbf{V}_i) = \mathcal{F}_{c,u,v}$ projecting the mesh vertices into image space by employing rasterization and mapping the image features at the 2D projected position (u,v) to the respective graph node. Note that the rasterization is occlusion-aware, *i.e.,* only visible vertices have an attached image feature. However, in a monocular setting, a significant amount of vertices is usually occluded although we also want to regress their deformations. Therefore, we argue that the whole image (of visible surface parts) can still guide the deformation state of occluded parts, e.g., the body pose can roughly give a hint of the deformations of occluded parts. Thus, we have a visibility sensitive feature attachment function $F_c(\mathbf{V}_i) = \mathcal{F}_{c,u,v}$ if $f_{\text{Visible},c}(\mathbf{V}_i)$ and $F_c(\mathbf{V}_i) = a(\mathcal{F}_c)$ otherwise, which assigns the projected feature if a vertex is visible and for occluded vertices, it assigns the average feature, *i.e.,* $a(\cdot)$ averages the per-pixel features over the spatial domain.

Visibility- and Rigidity-aware Graph Convolutional Network. Once each graph node has an image feature, we employ a graph CNN, *DGCN* [26] taking these per-node features and outputting the vertex displacement field $f_{\text{DGCN}}(F_c(\mathbf{V})) = \mathbf{D}' \in \mathbb{R}^{N\times 3}$. Here, the graph is defined by the template mesh itself. Nearly rigid human body parts (e.g., skin and shoes) should (if at all) only be coarsely deformed. Thus, we create a rigid mask $\mathbf{M} \in \mathbb{R}^{N\times 3}$ whose entries of row i are set to one if $r_i \leq \varepsilon_{\text{Rigid}}$ and zero otherwise. Here, $\varepsilon_{\text{Rigid}}$ is a threshold. The displacement field is defined as $\mathbf{D} = f_{\text{DGCN}}(F_c(\mathbf{V})) \circ \mathbf{M}$, where $\circ$ is the Hadamard product.

3.4 Training of EDefNet and DisplaceNet

Image-based Supervision. The silhouette loss $\mathcal{L}_{\text{sil}}$ encourages the deformed mesh to fit the image silhouettes from all the cameras. As such energy term can be stuck in local minima due to bad initialization, we employ a 2D multi-view landmark term $\mathcal{L}_{\text{mk}}$ as the difference between the projected 3D markers on the posed skeleton and the detected 2D markers on the multi-view images. However, the aforementioned losses are not sufficient to supervise fine deformations such as surface folds. Thus, we deploy a dense rendering loss, which takes the posed and deformed mesh and the static texture, renders it from various camera views, and compares the rendered images $R_c(\mathbf{V},\mathbf{L},\mathcal{T})$ under the camera's lighting condition $\mathbf{L}$ with the corresponding input frame $\mathcal{I}_c$, $\mathcal{L}_{\text{dr}}(\mathbf{V},L) = \sum_c \|R_c(\mathbf{V},\mathbf{L}_c,\mathcal{T}) - \mathcal{I}_c\|^2$. Assuming Lambertian surface and smooth lighting, we employ the spherical harmonics (*SH*) lighting model [53] to represent the scene lighting $L_c(\mathbf{V},\mathbf{l}_c)$ of each camera with 27 coefficients $\mathbf{l}_c \in \mathbb{R}^{9\times3}$. Then, the lighting condition for each camera can be computed as $L_c(\mathbf{V},\mathbf{l}_c) = \sum_{j=1}^{9} \mathbf{l}_{c,j} B_j(n_c(\mathbf{V}))$ where $n_c(\mathbf{V})$ is the pixel normals of the geometry from the camera c. Since scene lighting is assumed to be unknown, we also optimize it as described later.

Chamfer Loss. $\mathcal{L}_{\text{dr}}$ helps to recover in-camera-plane deformations but struggles with capturing deformations along the camera viewing direction. Therefore, we employ a Chamfer loss between the posed and deformed mesh $\mathbf{V}$ and the per-frame stereo reconstructions $\mathbf{V}_{\text{GT}}$: $\mathcal{L}_{\text{cf}}(\mathbf{V}) = \sum_i \min_j \|\mathbf{V}_i - \mathbf{V}_{\text{GT,j}}\|^2 + \sum_j \min_i \|\mathbf{V}_{\text{GT,j}} - \mathbf{V}_i\|^2$. Note that it can suffer from drifts along the surface, which $\mathcal{L}_{\text{dr}}$ can prevent. Thus, we use the combination of two.

Spatial Regularization. To regularize the deformations, we impose an as-rigid-as-possible regularizer on the deformation graph [66] and use material-aware weighting factors [25] to deal with different levels of rigidity. We also employ a Laplacian $\mathcal{L}_{\text{lap}}$ and isometry $\mathcal{L}_{\text{iso}}$ regularization on the deformed and posed template mesh similar to Habermann et al. [24].

Training stages. We train the *EDefNet* in two phases while keeping the trained *PoseNet* fixed. At this stage, the displacements are set to zero in the character representation. We first train the embedded deformation network using the combined loss $\mathcal{L}_{\text{EDefNet}} = \mathcal{L}_{\text{sil}} + \mathcal{L}_{\text{mk}} + \mathcal{L}_{\text{arap}}$ [66]. The different weights do not affect the training significantly; thus, we currently use an equal weighted sum in all experiments. Once converged, the lighting parameters are optimized as an in-between step. Note that this is only possible when coarse deformations are already learned, and thus the model already roughly overlays to the ground truth images. To optimize the lighting coefficients across all the frames in the training sequence, we minimize $\|R_c(\mathbf{V},L_c(\mathbf{V},\mathbf{l}_c),\mathcal{T}) - \mathcal{I}_c\|^2$ by iteratively sampling the training frames. Here, $\mathcal{T}$ is the static template texture. After convergence, we train EDefNet further and add the differentiable rendering loss [35, 47] $\mathcal{L}_{\text{dr}}$ using the optimized scene lighting and the Chamfer loss $\mathcal{L}_{\text{cf}}$.

While training *DisplaceNet*, the weights of *PoseNet* and *EDefNet* are fixed and the character representation adds the displacements on top of the embedded deformation. For supervision, we leverage the combined loss function $\mathcal{L}_{\text{DisplaceNet}} = \mathcal{L}_{\text{sil}} + \mathcal{L}_{\text{dr}} + \mathcal{L}_{\text{cf}} + \mathcal{L}_{\text{iso}} + \mathcal{L}_{\text{lap}}$. More details regarding the losses are provided in the supplemental document.

3.5 Tracking of Hands and Face

So far, only the skeletal pose and the surface deformations are tracked. To enable the joint tracking of hands and face as well, we replace the face and hand regions on the template mesh with a parametric 3D face model [7] and hand model [63] as described in the following.

Face. We leverage the parametric face model of Blanz and Vetter [7] with the surface geometry being defined as $\mathbf{V}_{\text{F}} = \overline{\mathbf{V}}_{\text{F}} + \sum_{i=1}^{80} \mathbf{w}_{\text{S},i} \sigma_{\text{S},i} \mathbf{B}_{\text{S},i} + \sum_{j=1}^{64} \mathbf{w}_{\text{E},j} \sigma_{\text{E},j} \mathbf{B}_{\text{E},j}$, where $\overline{\mathbf{V}}_{\text{F}}$ is the mean face. $\mathbf{B}_{\mathrm{S},i} \in \mathbb{R}^{80\times 53490}$, and $\mathbf{B}_{\mathrm{E},i} \in \mathbb{R}^{64\times 53490}$ are the PCA bases for shape and expression variations. $\sigma_{\mathrm{S},i}$ and $\sigma_{\mathrm{E},i}$ are the corresponding standard deviations. $\mathbf{w}_{\mathrm{S},i} \in \mathbb{R}^{80}$ and $\mathbf{w}_{\mathrm{E},i} \in \mathbb{R}^{64}$ are the face shape and expression parameters. We removed the neck and ear parts of the model to better fit our template in the following.

Hands. We utilize the parametric MANO model [63] for the hands embedded in a joint and fully articulated body and hand model called SMPL+H model defined as $\mathbf{M}_{\mathrm{SH}} = \overline{\mathbf{M}}_{\mathrm{SH}} + \sum_{i=1}^{16} \mathbf{w}_{\mathrm{SH},i}\mathbf{B}_{\mathrm{SH},i}$, where $\overline{\mathbf{M}}_{\mathrm{SH}}$ is the mean body shape with hands. $\mathbf{B}_{\mathrm{SH},i} \in \mathbb{R}^{16\times 6890}$ are the PCA basis for body shapes with hands and $\mathbf{w}_{\mathrm{SH},i} \in \mathbb{R}^{16}$ are the PCA coefficients. The posed and deformed mesh is defined as $\mathbf{V}_{\mathrm{SH}} = B(\mathbf{M}_{\mathrm{SH}}, \mathbf{W}, \theta_b, \theta_h)$ where $B(.)$ is the linear blend skinning function, $\mathbf{W}$ is the skinning weights, $\theta_b \in \mathbb{R}^{22}$, and $\theta_h \in \mathbb{R}^{15\times 2}$ are joint angles for body and hands. We set θ_b and θ_h to zero to obtain the deformed model in the canonical T-pose $\hat{\mathbf{V}}_{\mathrm{SH}}$ and then only consider the MANO vertices $\hat{\mathbf{V}}_{\mathrm{H}}$.

Unposing to the Canonical Pose. The original 3D template mesh $\mathcal{M}$ in the rigging pose can differ in terms of its local rigid rotation with respect to the face and hand model. For a better optimization for the personalized face model and hand stitching, we unpose the original 3D template to the canonical pose $\mathcal{M}_{\mathrm{tpose}}$ using Dual Quaternion Skinning.

Personalized Face Model. To retrieve the personalized face model, we fit the face model to the original template in the canonical pose $\mathcal{M}_{\mathrm{tpose}}$ as follows. First, we optimize the affine transform between the model and the template by optimizing the affine parameters including Euler angles for rotation $\alpha_{\mathrm{F}} \in \mathbb{R}^3$, translation vector $\mathbf{t}_{\mathrm{F}} \in \mathbb{R}^3$, and scaling $s_{\mathrm{F}} \in \mathbb{R}$. We convert the Euler angles α_{F} to a rotation matrix $\mathbf{R}_{\mathrm{F}} \in \mathbb{R}^{3\times 3}$. Then, the updated face model vertices $\mathbf{V}'_{\mathrm{F}}$ can be computed as $\mathbf{V}'_{\mathrm{F}} = s_{\mathrm{F}}\mathbf{R}_{\mathrm{F}}(\mathbf{V}_{\mathrm{F}}) + \mathbf{t}_{\mathrm{F}}$. To optimize the affine parameters, we manually mark 8 facial landmarks on the scanned template mesh and the face model (2 on each eye, 2 on lips, 1 on nose, 1 on jaw), respectively, and minimize the difference between the two sets in the least-squares sense. We then fix the affine transform and deform the face model to match the template geometry by optimizing the shape parameters $\mathbf{w}_{\mathrm{S}}$ and the expression parameters $\mathbf{w}_{\mathrm{E}}$. To this end, we minimize the Chamfer distance between the face model and template mesh as well as the distance between the two sets of markers. Then, we once more optimize the affine parameters by minimizing the aforementioned distances. Finally, we directly optimize the positions of the face model vertices by minimizing the Chamfer distance resulting in the updated face model position $\mathbf{V}''_{\mathrm{F}}$. We retrieve our final neutral face as $\hat{\mathbf{V}}_{\mathrm{F}} = \mathbf{V}''_{\mathrm{F}} - \sum_{j=1}^{64} \hat{w}_{\mathrm{E},j}\sigma_{\mathrm{E},j}\mathbf{b}_{\mathrm{E},j}$ where $\hat{w}_{\mathrm{E},j}$ are the optimized face expression parameters.

To connect the hand and face models with the template, we use an automated gap-filling technique in Blender [29] in the canonical space $\hat{\mathcal{M}}_{\mathrm{tpose}}$. Finally, we repose the template to get the updated template mesh in the rigging pose $\hat{\mathcal{M}}$. Therefore, the skinning weights of the closest template vertex are copied to the hand and face model vertices.

Regression of Hand and Face Parameters and Posing. Given an input frame f, we use the pre-trained model of Zhou et al. [97] to regress the facial expression parameters $\mathbf{w}_{\mathrm{E,f}}$ and the hand pose parameters $\theta_{H,f}$. Importantly, we do not leverage the regressed face shape, but we use our optimized face model as the identity. We then apply the regressed facial expression and hand pose parameters to the template in the canonical pose. Finally, we pose and deform the template by using the regressed embedded deformation $\mathbf{A}, \mathbf{T}$ from *EDefNet*, the displacement map $\mathbf{D}$ from *DisplaceNet*, and the body pose $\theta, \alpha, \mathbf{t}$ from *PoseNet*.

4 Results

Data. Our method is person-specific and we aim at generalization to novel poses and environments. Thus, we captured 3 subjects in different types of apparel (*e.g.,* skirts and

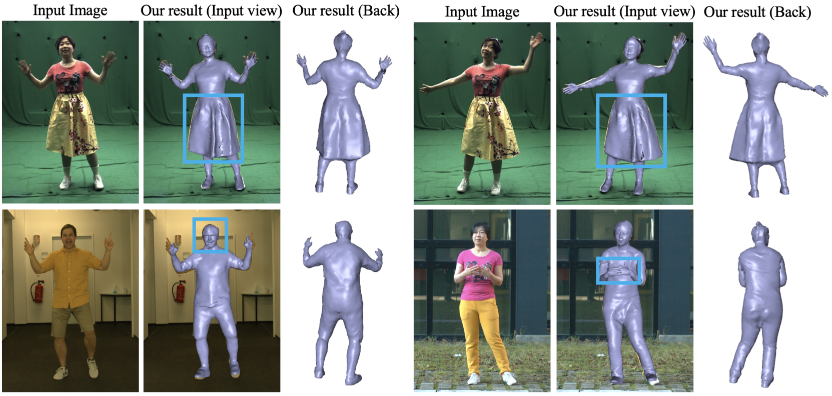

Figure 3: Qualitative results for subjects with different types of apparel, poses, and backgrounds. Our method not only precisely overlays onto the input images, but also captures the wrinkle patterns nicely. Even the occluded regions look plausible in the back views.

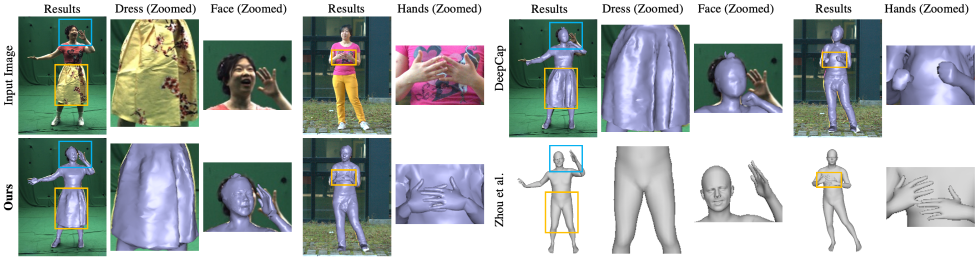

Figure 4: Comparisons. Our method can capture high-frequency details on the non-rigid clothing surfaces, facial expressions, and hand gestures. DeepCap [25] cannot dynamically capture face and hands, while Zhou et al. [97] cannot capture the non-rigid deformation. Our method can better capture dynamic non-rigid clothing details than DeepCap [25].

trousers). Per subject, we captured around 20k multi-view video frames for training (only in-studio green background). For testing, several separate 2k-frame videos in novel in-the-wild environments and poses are captured using a BlackMagic camera. We recorded in different environments (*e.g.,* indoors, outdoors, in-studio) to test the generalization of our approach to novel lighting conditions. All the sequences include a large variety of different and challenging motions. We apply a domain adaptation step proposed by Habermann et al. [25] by finetuning our pre-trained networks on the monocular test sequences. For quantitative evaluations, we also recorded 5 in-studio sequences to be able to acquire ground truth meshes using multi-view stereo [1].

Qualitative Results. We visualize monocular results in Fig. 3 and the supplement with different clothing, motions, and backgrounds. Our reconstruction jointly captures facial expressions, hand poses, and high-frequency details on clothing. It overlays precisely with the input images and achieves plausible results for the occluded areas. The recovered clothing wrinkles of the posed and deformed template match the ones in the input images.

Comparisons. There is no dataset with joint ground-truth skeletal pose, hands, face, and cloth tracking and obtaining such is far from being trivial. So it is hard to quantitatively evaluate them in our setting. As an alternative, we show extensive qualitative results for

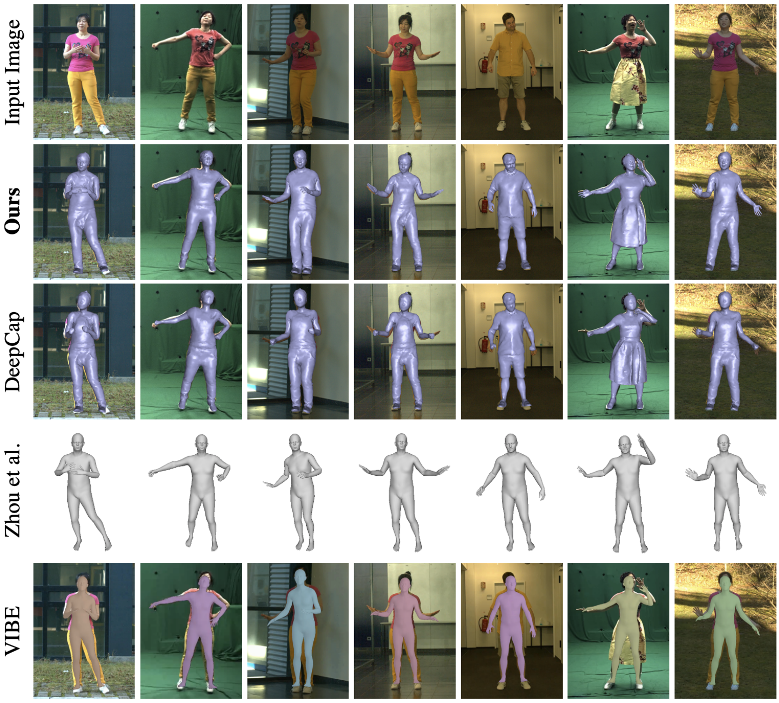

Figure 5: Qualitative comparison. Compared to other methods, HiFECap can better capture high-frequency details on the non-rigid clothing surfaces *and* facial expressions as well as hand gestures and be generalizable to different clothing, motions, and backgrounds.

face and hands in Fig. 4 and our supplemental material. In Fig. 5, we further compare our results qualitatively with related approaches. Compared to our approach, DeepCap [25] cannot capture high-frequency details on the clothing due to the limited capacity of the embedded graph and the silhouette-only supervision strategy. Our method can capture more accurate clothing details than DeepCap corresponding to the input video. Although DeepCap tracks the clothing, it outputs very different (mostly coarse and global) wrinkles compared to the ones observed in the input. By leveraging image convolutions and graph convolutions, our new architecture, DisplaceNet, regresses the per-vertex displacement field. It captures the dynamic high-frequency details on the nonrigid deforming surface while DeepCap only deforms the static clothing of the input template by matching the silhouette of the deforming clothing using 2D supervision. Furthermore, DeepCap can only capture the pose and clothing deformations, while our approach can dynamically capture facial expressions and hand poses. Our input video captures the entire body without additional information about hands, and we further localize the hands to capture these parts. Thus, existing hand-only methods cannot be directly applied to our setting as our method requires cropping and alignments of the human's hands, face, and body. Concerning full-body methods, we show superior hands and face capture results compared to previous work. Zhou et al. [96] jointly regresses facial expressions, hand poses, and the body pose, but it is not able to capture the non-rigid deformation of the clothing at all, and VIBE [43] only captures body and hand poses.

Quantitative Results. We evaluate the accuracy of the recovered non-rigid deformation between our method and related works [25, 43, 97]. Tab. 4 shows the quantitative results of these methods for three test sequences by computing the average Chamfer distance and the

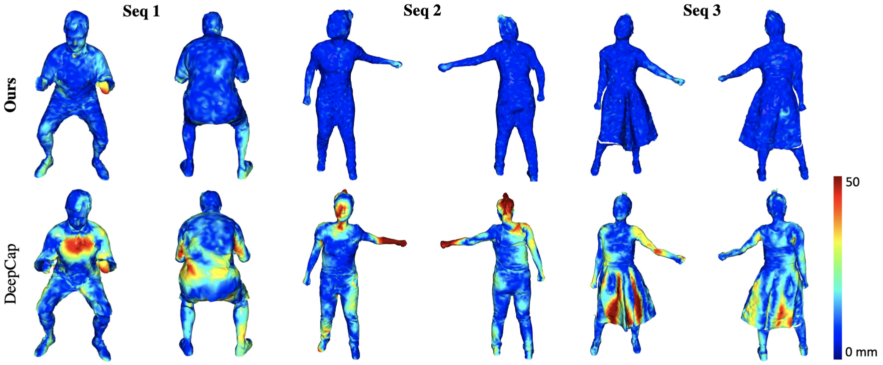

Figure 6: Visualization of the per-vertex error (MSE) of our method and DeepCap [25] compared to the ground truth meshes. Our method has significantly lower error indicating our method better captures high-frequency non-rigid surface deformations.

	Sequence 1		Sequence 2		Sequence 3	
Method	**Chamfer↓**	**Hausdorff↓**	**Chamfer↓**	**Hausdorff↓**	**Chamfer↓**	**Hausdorff↓**
Ours	**7.25**	**39.72**	**9.21**	**40.07**	**7.26**	**32.69**
DeepCap [25]	21.09	77.83	14.32	107.21	17.88	98.49
Zhou et al [97]	24.49	133.95	51.35	230.42	34.83	157.11
VIBE [43]	47.21	121.15	72.00	229.90	82.24	224.60

Table 1: Quantitative comparisons. Our method significantly outperforms other approaches in terms of Chamfer distance and Hausdorff distance with respect to the ground truth. The accuracy of our method increased by almost 50% compared to the state-of-the-art DeepCap. Our approach can better capture high-frequency details on the dynamic non-rigid surfaces.

average symmetric Hausdorff distance between the output and ground truth meshes. We observe that our method achieves higher accuracy in terms of both metrics confirming that our reconstruction results can capture high-frequency details on the non-rigid parts. In Fig. 6, we show the 3D reconstruction results of our approach and the state-of-the-art approach DeepCap and visualize their per-vertex errors to the ground truth meshes. Note that we do not apply the dynamic hand pose and facial expressions here to evaluate the non-rigid clothing deformations separately. Our method especially outperforms DeepCap in the dynamic clothing areas indicating that such dynamic deformations are better recovered by our approach.

5 Conclusion

In this paper, we presented HiFECap, the first monocular human performance capture approach, which jointly tracks the body pose, hand gestures, facial expressions, and high-fidelity non-rigid surface deformations. We showed that higher-fidelity character surface tracking can be achieved by adding a dedicated displacement network to the character deformation process, which is a hybrid network architecture leveraging image convolutions and graph convolutions with locality preserving receptive fields. Further, tightly coupling the template with parametric hand and face models enables the tracking of all aspects of the human. In our experiments, we validated these design choices and show that we improve the current state-of-the-art in terms of space-time coherent surface tracking. While this work is a clear step towards expressive capture of humans, we still believe that there is a lot of future work to be done, especially in the areas of physically correct human tracking and real-time performance, which would enable monocular performance capture in VR and AR settings.

6 Acknowledgements

The authors from MPII were supported by the ERC Consolidator Grant 4DRepLy (770784). We would like to thank Mengyu Chu for her participation in our experiments and Zhi Li for the voice-over of our video.

References

[1] Agisoft metashape. https://www.agisoft.com/.

[2] Benjamin Allain, Jean-Sébastien Franco, and Edmond Boyer. An Efficient Volumetric Framework for Shape Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[3] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[4] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. SCAPE: Shape Completion and Animation of People. *ACM Transactions on Graphics*, 2005.

[5] Alexandru O Bălan and Michael J Black. The naked truth: Estimating body shape under clothing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2008.

[6] Alexandru O Balan, Leonid Sigal, Michael J Black, James E Davis, and Horst W Haussecker. Detailed human shape and pose from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[7] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 1999.

[8] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[9] Gunilla Borgefors. Distance transformations in digital images. *Computer Vision, Graphics, and Image Processing*, 1986.

[10] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[11] Matthieu Bray, Pushmeet Kohli, and Philip HS Torr. Posecut: Simultaneous segmentation and 3d pose estimation of humans using dynamic graph-cuts. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2006.

[12] Thomas Brox, Bodo Rosenhahn, Daniel Cremers, and Hans-Peter Seidel. High accuracy optical flow serves 3-d pose tracking: exploiting contour and flow based constraints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2006.

[13] Thomas Brox, Bodo Rosenhahn, Juergen Gall, and Daniel Cremers. Combined region and motion-based 3d tracking of rigid and articulated objects. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 2009.

[14] Cedric Cagniart, Edmond Boyer, and Slobodan Ilic. Free-form mesh tracking: a patch-based approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[15] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[16] Joel Carranza, Christian Theobalt, Marcus A. Magnor, and Hans-Peter Seidel. Free-viewpoint video of human actors. *ACM Transactions on Graphics*, 2003.

[17] Xingyu Chen, Yufeng Liu, Chongyang Ma, Jianlong Chang, Huayan Wang, Tian Chen, Xiaoyan Guo, Pengfei Wan, and Wen Zheng. Camera-space hand mesh recovery via semantic aggregation and adaptive 2d-1d registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[18] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics*, 2015.

[19] Edilson De Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. Performance capture from sparse multi-view video. In *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 2008.

[20] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael Black. Collaborative regression of expressive bodies using moderation. In *Proceedings of International Conference on 3D Vision (3DV)*, 2021.

[21] 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 Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[22] Juergen Gall, Carsten Stoll, Edilson De Aguiar, Christian Theobalt, Bodo Rosenhahn, and Hans-Peter Seidel. Motion capture using joint skeleton tracking and surface estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[23] Liuhao 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[24] Marc Habermann, Weipeng Xu, Michael Zollhöfer, Gerard Pons-Moll, and Christian Theobalt. Livecap: Real-time human performance capture from monocular video. *ACM Transactions on Graphics*, 2019.

[25] Marc Habermann, Weipeng Xu, Michael Zollhoefer, 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 (CVPR)*, 2020.

[26] Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Real-time deep dynamic characters. In *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 2021.

[27] Nils Hasler, Hanno Ackermann, Bodo Rosenhahn, Thorsten Thormählen, and Hans-Peter Seidel. Multilinear pose and body shape estimation of dressed subjects from image sets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[29] Roland Hess. *Blender Foundations: The Essential Guide to Learning Blender 2.6*. Focal Press, 2010. ISBN 0240814304, 9780240814308.

[30] Nikolas Hesse, Sergi Pujades, Javier Romero, Michael J. Black, Christoph Bodensteiner, Michael Arens, Ulrich G. Hofmann, Uta Tacke, Mijna Hadders-Algra, Raphael Weinberger, Wolfgang Muller-Felber, and A. Sebastian Schroeder. Learning an infant body model from RGB-D data for accurate full body motion analysis. In *Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, September 2018.

[31] C.-H. Huang, B. Allain, J.-S. Franco, N. Navab, S. Ilic, and E. Boyer. Volumetric 3d tracking by detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[32] Yinghao Huang, Federica Bogo, Christoph Lassner, Angjoo Kanazawa, Peter V. Gehler, Javier Romero, Ijaz Akhter, and Michael J. Black. Towards accurate markerless human shape and pose estimation over time. In *Proceedings of International Conference on 3D Vision (3DV)*, 2017.

[33] Zeng Huang, Tianye Li, Weikai Chen, Yajie Zhao, Jun Xing, Chloe LeGendre, Linjie Luo, Chongyang Ma, and Hao Li. Deep Volumetric Video From Very Sparse Multi-View Performance Capture. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[34] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[35] Yue Jiang, Dantong Ji, Zhizhong Han, and Matthias Zwicker. Sdfdiff: Differentiable rendering of signed distance fields for 3d shape optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[36] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[37] Petr Kadlecek, Alexandru-Eugen Ichim, Tiantian Liu, Jaroslav Krivanek, and Ladislav Kavan. Reconstructing personalized anatomical models for physics-based body animation. *ACM Transactions on Graphics*, 35(6), 2016.

[38] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[39] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[40] Ladislav Kavan, Steven Collins, Jiří Žára, and Carol O'Sullivan. Skinning with dual quaternions. In *Proceedings of the Symposium on Interactive 3D Graphics and Games (I3D)*, 2007.

[41] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Pérez, Christian Richardt, Michael Zollöfer, and Christian Theobalt. Deep video portraits. *ACM Transactions on Graphics*, 2018.

[42] Meekyoung Kim, Gerard Pons-Moll, Sergi Pujades, Sungbae Bang, Jinwwok Kim, Michael Black, and Sung-Hee Lee. Data-driven physics for human soft tissue animation. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 2017.

[43] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[44] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[45] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V.Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[46] Yue Li, Marc Habermann, Bernhard Thomaszewski, Stelian Coros, Thabo Beeler, and Christian Theobalt. Deep physics-aware inference of cloth deformation for monocular human performance capture. In *Proceedings of International Conference on 3D Vision (3DV)*, 2021.

[47] Shichen Liu, Weikai Chen, Tianye Li, and Hao Li. Soft rasterizer: Differentiable rendering for unsupervised single-view mesh reconstruction. *arXiv preprint arXiv:1901.05567*, 2019.

[48] Yebin Liu, Carsten Stoll, Juergen Gall, Hans-Peter Seidel, and Christian Theobalt. Markerless motion capture of interacting characters using multi-view image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[49] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 2015.

[50] Wojciech Matusik, Chris Buehler, Ramesh Raskar, Steven J Gortler, and Leonard McMillan. Image-based visual hulls. In *ACM Transactions on Graphicss*, 2000.

[51] 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. 2017.

[52] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[53] Claus Müller. *Spherical harmonics*. 2006.

[54] Armin Mustafa, Hansung Kim, Jean-Yves Guillemaut, and Adrian Hilton. General dynamic scene reconstruction from multiple view video. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[55] Sang Il Park and Jessica K Hodgins. Data-driven modeling of skin and muscle deformation. In *ACM Transactions on Graphics*, 2008.

[56] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3d human pose and shape from a single color image. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[57] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[58] Gerard Pons-Moll, Javier Romero, Naureen Mahmood, and Michael J Black. Dyna: A model of dynamic human shape in motion. *ACM Transactions on Graphics*, 2015.

[59] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics*, 2017.

[60] Alin-Ionut Popa, Mihai Zanfir, and Cristian Sminchisescu. Deep multitask architecture for integrated 2d and 3d human sensing. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[61] Helge Rhodin, Nadia Robertini, Dan Casas, Christian Richardt, Hans-Peter Seidel, and Christian Theobalt. General automatic human shape and motion capture using volumetric contour cues. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[62] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net: Localization-classification-regression for human pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[63] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 2017.

[64] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, 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 (ICCV)*, 2019.

[65] 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 (CVPR)*, 2020.

[66] Olga Sorkine and Marc Alexa. As-rigid-as-possible surface modeling. In *Proceedings of the Fifth Eurographics Symposium on Geometry Processing (SGP)*, 2007.

[67] Jonathan Starck and Adrian Hilton. Surface capture for performance-based animation. *IEEE Computer Graphics and Applications*, 2007.

[68] Robert W. Sumner, Johannes Schmid, and Mark Pauly. Embedded deformation for shape manipulation. *ACM Transactions on Graphics*, 2007.

[69] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[70] Bugra Tekin, Pablo Márquez-Neila, Mathieu Salzmann, and Pascal Fua. Learning to fuse 2d and 3d image cues for monocular body pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[71] Ayush Tewari, Michael Zollöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Theobalt Christian. MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[72] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[73] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[74] Denis Tome, Chris Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[75] Treedys. Treedys. https://www.treedys.com/, 2020.

[76] Xiaoguang Tu, Jian Zhao, Zihang Jiang, Yao Luo, Mei Xie, Yang Zhao, Linxiao He, Zheng Ma, and Jiashi Feng. Joint 3d face reconstruction and dense face alignment from a single image with 2d-assisted self-supervised learning. *arXiv preprint arXiv:1903.09359*, 2019.

[77] Gül Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[78] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. Bodynet: Volumetric inference of 3d human body shapes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[79] Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović. Articulated mesh animation from multi-view silhouettes. In *ACM Transactions on Graphics, (Proc. SIGGRAPH)*. 2008.

[80] Daniel Vlasic, Pieter Peers, Ilya Baran, Paul Debevec, Jovan Popović, Szymon Rusinkiewicz, and Wojciech Matusik. Dynamic shape capture using multi-view photometric stereo. In *ACM Transactions on Graphics, (Proc. SIGGRAPH)*. 2009.

[81] Jiayi Wang, Franziska Mueller, Florian Bernard, Suzanne Sorli, Oleksandr Sotnychenko, Neng Qian, Miguel A. Otaduy, Dan Casas, and Christian Theobalt. RGB2Hands: Real-Time Tracking of 3D Hand Interactions from Monocular RGB Video. *ACM Transactions on Graphics*, 2020.

[82] Michael Waschbüsch, Stephan Würmlin, Daniel Cotting, Filip Sadlo, and Markus Gross. Scalable 3d video of dynamic scenes. *The Visual Computer*, 2005.

[83] Chenglei Wu, Kiran Varanasi, and Christian Theobalt. Full body performance capture under uncontrolled and varying illumination: A shading-based approach. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012.

[84] Chenglei Wu, Carsten Stoll, Levi Valgaerts, and Christian Theobalt. On-set performance capture of multiple actors with a stereo camera. *ACM Transactions on Graphics*, 2013.

[85] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[86] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[87] Donglai Xiang, Fabian Prada, Chenglei Wu, and Jessica Hodgins. Monoclothcap: Towards temporally coherent clothing capture from monocular rgb video. In *Proceedings of International Conference on 3D Vision (3DV)*, 2020.

[88] Lan Xu, Weipeng Xu, Vladislav Golyanik, Marc Habermann, Lu Fang, and Christian Theobalt. Eventcap: Monocular 3d capture of high-speed human motions using an event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[89] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. Monoperfcap: Human performance capture from monocular video. *ACM Transactions on Graphics*, 2018.

[90] Jinlong Yang, Jean-Sébastien Franco, Franck Hétroy-Wheeler, and Stefanie Wuhrer. Estimation of Human Body Shape in Motion with Wide Clothing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, Amsterdam, Netherlands, 2016.

[91] Chao Zhang, Sergi Pujades, Michael Black, and Gerard Pons-Moll. Detailed, accurate, human shape estimation from clothed 3D scan sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[92] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[93] Xiong Zhang, Hongsheng Huang, Jianchao Tan, Hongmin Xu, Cheng Yang, Guozhu Peng, Lei Wang, and Ji Liu. Hand image understanding via deep multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[94] 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 (ICCV)*, 2019.

[95] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: A weakly-supervised approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[96] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[97] Yuxiao Zhou, Marc Habermann, Ikhsanul Habibie, Ayush Tewari, Christian Theobalt, and Feng Xu. Monocular real-time full body capture with inter-part correlations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.