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

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 [52, 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, 31, 50, 61, 67, 80, 82]. Prior works reconstruct surface deformations based on person-specific template meshes [14, 16, 19] or a volumetric representation [6]. 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, 58]. Approaches relying on them often ignore clothing by treating it as noise [4], or estimate only naked body shape [6, 32, 90, 91]. Some techniques track facial expressions [41] and hands [52, 56] 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 lever-
ages 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 [24], and can also jointly capture body pose with facial expressions and hand gestures [20, 57, 86, 97]. PIFuHD [64] and SelfRecon [34] work for standing poses and do not generalize to arbitrary poses. Moreover, PIFu[HD] [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).
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 $V_{GT,f}$ for each frame. As input to our method, we randomly sample cropped and segmented frames $I_{f,c}$ 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 $t \in \mathbb{R}^3$, 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 $V_i$ to account for the different deformation properties of varying materials. We further define a downsampled version of the mesh as the underlying embedded graph $G$ and model deformations from coarse to fine using an embedded deformation [68]. $G$ is parameterized with $A \in \mathbb{R}^{K \times 3}$ and $T \in \mathbb{R}^{K \times 3}$ 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 $D$, i.e., we assign a displacement vector $d_i \in \mathbb{R}^3$ 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 [26], PoseNet and \( EDefNet \) are both Resnet50-based networks [26] that take the image \( I_c \) of view \( c \) as input. The network architecture of \( EDefNet \) is the same as \( DefNet \) in DeepCap [26], 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 [26]. We supervise PoseNet with a multi-view 2D keypoint loss and a joint angle regularizer as proposed in DeepCap [26]. \( EDefNet \) 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, \( \text{DisplaceNet} \), which takes the input image \( I_c \) 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, \( \text{e.g.} \), wrinkles from the input image. More precisely, it takes the input frame \( I_c \in \mathbb{R}^{256 \times 256 \times 3} \) and computes a latent feature map \( f_{DUNet,c}(I_c) = \mathcal{F}_c \in \mathbb{R}^{256 \times 256 \times 32} \) 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 \( P_c(V_i) = \mathcal{F}_{c,u,v} \) 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, \( \text{i.e.} \), 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, \( \text{e.g.} \), the body pose can roughly give a hint of the deformations of occluded parts. Thus, we have a visibility sensitive feature attachment function \( F_c(V_i) = \mathcal{F}_{c,u,v} \) if \( f_{\text{visible},c}(V_i) \) and \( F_c(V_i) = a(\mathcal{F}_c) \) otherwise, which assigns the projected feature if a vertex is visible and for occluded vertices, it assigns the average feature, \( \text{i.e.} \), \( a(\cdot) \) 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, \( \text{DGCN} \) [26] taking these per-node features and outputting the vertex displacement field \( f_{\text{DGCN}}(F_c(V)) = \mathbf{D}' \in \mathbb{R}^{N \times 3} \). Here, the graph is defined by the template mesh itself. Nearly rigid human body parts (\( \text{e.g.} \), 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 \( \mathbf{D} = f_{\text{DGCN}}(F_c(V)) \circ \mathbf{M} \), 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 $R_c(V, L, T)$ under the camera’s lighting condition $L$ with the corresponding input frame $I_c$, $\mathcal{L}_{\text{dr}}(V, L) = \sum_c \|R_c(V, L_c, T) - I_c\|^2$. Assuming Lambertian surface and smooth lighting, we employ the spherical harmonics (SH) lighting model \cite{Lafortune20} to represent the scene lighting $L_c(V, l_c)$ of each camera with 27 coefficients $l_c \in \mathbb{R}^{9 \times 3}$. Then, the lighting condition for each camera can be computed as $L_c(V, l_c) = \sum_{j=1}^{9} l_{c,j} B_j(n_c(V))$ where $n_c(V)$ 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 $V$ and the per-frame stereo reconstructions $V_{\text{GT}}$: $\mathcal{L}_{\text{cf}}(V) = \sum_i \min_j \|V_i - V_{\text{GT},j}\|^2 + \sum_j \min_i \|V_{\text{GT},j} - V_i\|^2$. 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 \cite{Schaefer2015} and use material-aware weighting factors \cite{Yang19} 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. \cite{Habermann20}.

**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 $\mathcal{L}_{\text{EDefNet}} = \mathcal{L}_{\text{sil}} + \mathcal{L}_{\text{mk}} + \mathcal{L}_{\text{arap}}$ \cite{Schaefer2015}. 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 $\|R_c(V, L_c(V, l_c), T) - I_c\|^2$ by iteratively sampling the training frames. Here, $T$ is the static template texture. After convergence, we train EDefNet further and add the differentiable rendering loss \cite{Yang19,Schaefer2015} $\mathcal{L}_{\text{dr}}$ 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 \cite{Blanz99} and hand model \cite{Wu15} as described in the following.

**Face.** We leverage the parametric face model of Blanz and Vetter \cite{Blanz99} with the surface geometry being defined as $V_F = \overline{V}_F + \sum_{i=1}^{80} w_{S, i} \sigma_{S, i} S_{B} + \sum_{j=1}^{64} w_{E, j} \sigma_{E, j} E_{B, j}$, where $\overline{V}_F$ is the
mean face. \( \mathbf{B}_{S,i} \in \mathbb{R}^{80 \times 53490} \), and \( \mathbf{B}_{E,i} \in \mathbb{R}^{64 \times 53490} \) are the PCA bases for shape and expression variations. \( \sigma_{S,i} \) and \( \sigma_{E,i} \) are the corresponding standard deviations. \( \mathbf{w}_{S,i} \in \mathbb{R}^{80} \) and \( \mathbf{w}_{E,i} \in \mathbb{R}^{64} \) 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 \( \mathbf{M}_{SH} = \mathbf{M}_{SH} + \sum_{i=1}^{16} \mathbf{w}_{SH,i} \mathbf{B}_{SH,i} \), where \( \mathbf{M}_{SH} \) is the mean body shape with hands. \( \mathbf{B}_{SH,i} \in \mathbb{R}^{16 \times 6890} \) are the PCA basis for body shapes with hands and \( \mathbf{w}_{SH,i} \in \mathbb{R}^{16} \) are the PCA coefficients. The posed and deformed mesh is defined as \( \mathbf{V}_{SH} = B(\mathbf{M}_{SH}, \mathbf{W}, \theta_b, \theta_h) \) where \( B(\cdot) \) 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 \( \theta_b \) and \( \theta_h \) to zero to obtain the deformed model in the canonical T-pose \( \hat{\mathbf{V}}_{SH} \) and then only consider the MANO vertices \( \hat{\mathbf{V}}_H \).

**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 \( \mathcal{M}_{tpose} \) 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 \( \mathcal{M}_{tpose} \) as follows. First, we optimize the affine transform between the model and the template by optimizing the affine parameters including Euler angles for rotation \( \alpha_F \in \mathbb{R}^3 \), translation vector \( \mathbf{t}_F \in \mathbb{R}^3 \), and scaling \( \mathbf{s}_F \in \mathbb{R} \). We convert the Euler angles \( \alpha_F \) to a rotation matrix \( \mathbf{R}_F \in \mathbb{R}^{3 \times 3} \). Then, the updated face model vertices \( \mathbf{V}_F' \) can be computed as \( \mathbf{V}_F' = \mathbf{s}_F \mathbf{R}_F(\mathbf{V}_F) + \mathbf{t}_F \). 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 \( \mathbf{w}_S \) and the expression parameters \( \mathbf{w}_E \). 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 \( \mathbf{V}_F'' \). We retrieve our final neutral face as \( \hat{\mathbf{V}}_F = \mathbf{V}_F'' - \sum_{j=1}^{64} \hat{\mathbf{w}}_{E,j} \sigma_{E,j} \mathbf{b}_{E,j} \) where \( \hat{\mathbf{w}}_{E,j} \) 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 [20] in the canonical space \( \hat{\mathcal{M}}_{tpose} \). Finally, we repose the template to get the updated template mesh in the rigging pose \( \mathcal{M} \). 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 \( \mathbf{w}_{E,f} \) 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
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.

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Sequence 1</th>
<th>Sequence 2</th>
<th>Sequence 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Chamfer↓</td>
<td>Hausdorff↓</td>
<td>Chamfer↓</td>
</tr>
<tr>
<td>Ours</td>
<td>7.25</td>
<td>39.72</td>
<td>9.21</td>
</tr>
<tr>
<td>DeepCap [25]</td>
<td>21.09</td>
<td>77.83</td>
<td>14.32</td>
</tr>
<tr>
<td>Zhou et al [97]</td>
<td>24.49</td>
<td>133.95</td>
<td>51.35</td>
</tr>
<tr>
<td>VIBE [43]</td>
<td>47.21</td>
<td>121.15</td>
<td>72.00</td>
</tr>
</tbody>
</table>

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


