DeepCap: Monocular Human Performance Capture Using Weak Supervision

Marc Habermann\textsuperscript{1,2} Weipeng Xu\textsuperscript{1,2} Michael Zollhoefer\textsuperscript{3} Gerard Pons-Moll\textsuperscript{1,2} Christian Theobalt\textsuperscript{1,2}

\textsuperscript{1}Max Planck Institute for Informatics, \textsuperscript{2}Saarland Informatics Campus, \textsuperscript{3}Stanford University

Abstract

Human performance capture is a highly important computer vision problem with many applications in movie production and virtual/augmented reality. Many previous performance capture approaches either required expensive multi-view setups or did not recover dense space-time coherent geometry with frame-to-frame correspondences. We propose a novel deep learning approach for monocular dense human performance capture. Our method is trained in a weakly supervised manner based on multi-view supervision completely removing the need for training data with 3D ground truth annotations. The network architecture is based on two separate networks that disentangle the task into a pose estimation and a non-rigid surface deformation step. Extensive qualitative and quantitative evaluations show that our approach outperforms the state of the art in terms of quality and robustness.

1. Introduction

Human performance capture, i.e., the space-time coherent 4D capture of full pose and non-rigid surface deformation of people in general clothing, revolutionized the film and gaming industry in recent years. Apart from visual effects, it has many use cases in generating personalized dynamic virtual avatars for telepresence, virtual try-on, mixed reality, and many other areas. In particular for the latter applications, being able to perform capture humans from monocular video would be a game changer. The majority of established monocular methods only captures articulated motion (including hands or sparse facial expression at most). However, the monocular tracking of dense full-body deformations of skin and clothing, in addition to articulated pose, which play an important role in producing realistic virtual characters, is still at its infancy.

In literature, multi-view marker-less methods [13, 14, 15, 17, 24, 29, 50, 55, 81, 82, 86, 64, 65] have shown compelling results. However, these approaches rely on well-controlled multi-camera studios (typically with green screen), which prohibits them from being used for location shootings of films and telepresence in living spaces.

Recent monocular human modeling approaches have shown compelling reconstructions of humans, including clothing, hair and facial details [70, 99, 2, 3, 9, 60, 52]. Some directly regress voxels [28, 99] or the continuous occupancy of the surface [70]. Since predictions are pixel aligned, reconstructions have nice detail, but limbs are often missing, especially for difficult poses. Moreover, the recovered motion is not factorized into articulation and non-rigid deformation, which prevents the computer-graphics style control over the reconstructions that is required in many of the aforementioned applications. Importantly, surface vertices are not tracked over time, so no space-time coherent model is captured. Another line of work predicts deformations or displacements to an articulated template, which prevents missing limbs and allows more control [2, 9, 5, 67]. However, these works do not capture motion and the surface deformations.

The state-of-the-art monocular human performance capture methods [89, 32] densely track the deformation of the surface. They leverage deep learning-based sparse keypoint detections and perform an expensive template fitting afterwards. In consequence, they can only non-rigidly fit to the input view and suffer from instability. By contrast, we present the first learning-based method that jointly infers the articulated and non-rigid 3D deformation parameters in a single feed-forward pass at much higher performance, accuracy and robustness. The core of our method is a CNN model which integrates a fully differentiable mesh template parameterized with pose and an embedded deformation graph. From a single image, our network predicts the skeletal pose, and the rotation and translation parameters for each node in the deformation graph. In stark contrast to implicit representations [70, 99, 22], our mesh-based

Figure 1. We present the first learning-based approach for dense monocular human performance capture using weak multi-view supervision that not only predicts the pose but also the space-time coherent non-rigid deformations of the model surface.
method tracks the surface vertices over time, which is crucial for adding semantics, and for texturing and rendering in graphics. Further, by virtue of our parameterization, our model always produces a human surface without missing limbs, even during occlusions and out-of-plane motions.

While previous methods [70, 99, 2, 9] rely on 3D ground truth for training, our method is weakly supervised from multi-view images. To this end, we propose a fully differentiable architecture which is trained in an analysis-by-synthesis fashion, without explicitly using any 3D ground truth annotation. Specifically, during training, our method only requires a personalized template mesh of the actor and a multi-view video sequence of the actor performing various motions. Then, our network learns to predict 3D pose and dense non-rigidly deformed surface shape by comparing its single image feed-forward predictions in a differentiable manner against the multi-view 2D observations. At test time, our method only requires a single-view image as input and produces a deformed template matching the actor’s non-rigid motion in the image. In summary, the main technical contributions of our work are:

- A learning-based 3D human performance capture approach that jointly tracks the skeletal pose and the non-rigid surface deformations from monocular images.
- A new differentiable representation of deforming human surfaces which enables training from multi-view video footage directly.

Our new model achieves high quality dense human performance capture results on our new challenging dataset, demonstrating, qualitatively and quantitatively, the advantages of our approach over previous work. We experimentally show that our method produces reconstructions of higher accuracy and 3D stability, in particular in depth, than related work, also under difficult poses.

2. Related Work

In the following, we focus on related work in the field of dense 3D human performance capture and do not review work on sparse 2D pose estimation.

**Capture using Parametric Models.** Monocular human performance capture is an ill-posed problem due to its high dimensionality and ambiguity. Low-dimensional parametric models can be employed as shape and deformation prior. First, model-based approaches leverage a set of simple geometric primitives [63, 74, 71, 54]. Recent methods employ detailed statistical models learned from thousands of high-quality 3D scans [6, 33, 59, 65, 51, 41, 45, 85, 35, 97, 10]. Deep learning is widely used to obtain 2D and/or 3D joint detections or 3D vertex positions that can be used to inform model fitting [37, 48, 53, 11, 46]. An alternative is to regress model parameters directly [42, 62, 43]. Beyond body shape and pose, recent models also include facial expressions and hand motion [61, 88, 40, 69] leading to very expressive reconstruction results. Since parametric body models do not represent garments, variation in clothing cannot be reconstructed, and therefore many methods recover the naked body shape under clothing [8, 7, 95, 90]. The full geometry of the actor can be reconstructed by non-rigidly deforming the base parametric model to better fit the observations [68, 3, 4]. But they can only model tight clothes such as T-shirts and pants, but not loose apparel which has a different topology than the body model, such as skirts. To overcome this problem, ClothCap [64] captures the body and clothing separately, but requires active multi-view setups. Physics based simulations have recently been leveraged to constrain tracking (SimulCap [78]), or to learn a model of clothing on top of SMPL (TailorNet [60]).

**Depth-based Template-free Capture.** Most approaches based on parametric models ignore clothing. The other side of the spectrum are prior-free approaches based on one or multiple depth sensors. Capturing general non-rigidly deforming scenes [73, 31], even at real-time frame rates [57, 39, 31], is feasible, but only works reliably for small, controlled, and slow motions. Higher robustness can be achieved by using higher frame rate sensors [30, 47] or multi-view setups [91, 27, 58, 26, 96]. Techniques that are specifically tailored to humans increase robustness [93, 94, 92] by integrating a skeletal motion prior [93] or a parametric model [94, 84]. HybridFusion [98] additionally incorporates a sparse set of inertial measurement units. These fusion-style volumetric capture techniques [36, 1, 49, 23, 66] achieve impressive results, but do not establish a set of dense correspondences between all frames. In addition, such depth-based methods do not directly generalize to our monocular setting, have a high power consumption, and typically do not work well under sunlight.

**Monocular Template-free Capture.** Quite recently, fueled by the progress in deep learning, many template-free monocular reconstruction approaches have been proposed. Due to their regular structure, many implicit reconstruction techniques [80, 99] make use of uniform voxel grids. DeepHum [99] combines a coarse scale volumetric reconstruction with a refinement network to add high-frequency details. Multi-view CNNs can map 2D images to 3D volumetric fields enabling reconstruction of a clothed human body at arbitrary resolution [38]. SiCloPe [56] reconstructs a complete textured 3D model, including cloth, from a single image. PiFu [70] regresses an implicit surface representation that locally aligns pixels with the global context of the corresponding 3D object. Unlike voxel-based representations, this implicit per-pixel representation is more memory.
efficient. These approaches have not been demonstrated to generalize well to strong articulation. Furthermore, implicit approaches do not recover frame-to-frame correspondences which are of paramount importance for downstream applications, e.g., in augmented reality and video editing. In contrast, our method is based on a mesh representation and can explicitly obtain the per-vertex correspondences over time while being slightly less general.

**Template-based Capture.** An interesting trade-off between being template-free and relying on parametric models are approaches that only employ a template mesh as prior. Historically, template-based human performance capture techniques exploit multi-view geometry to track the motion of a person [76]. Some systems also jointly reconstruct and obtain a foreground segmentation [13, 15, 50, 87]. Given a sufficient number of multi-view images as input, some approaches [21, 17, 24] align a personalized template model to the observations using non-rigid registration. All the aforementioned methods require expensive multi-view setups and are not practical for consumer use. Depth-based techniques enable template tracking from less cameras [100, 91] and reduced motion models [86, 29, 81, 50] increase tracking robustness. Recently, capturing 3D dense human body deformation just with a single RGB camera has been enabled [89] and real-time performance has been achieved [32]. However, their methods rely on expensive optimization leading either to very long per-frame computation times [89] or the need for two graphics cards [32]. Similar to them, our approach also employs a person-specific template mesh. But differently, our method directly learns to predict the skeletal pose and the non-rigid surface deformations. As shown by our experimental results, benefiting from our multi-view based self-supervision, our reconstruction accuracy significantly outperforms the existing methods.

### 3. Method

Given a single RGB video of a moving human in general clothing, our goal is to capture the dense deforming surface of the full body. This is achieved by training a neural network consisting of two components: As illustrated in Fig. 2, our pose network, *PoseNet*, estimates the skeletal pose of the actor in the form of joint angles from a monocular image (Sec. 3.2). Next, our deformation network, *DefNet*, regresses the non-rigid deformation of the dense surface, which cannot be modeled by the skeletal motion, in the embedded deformation graph representation (Sec. 3.3). To avoid generating dense 3D ground truth annotation, our network is trained in a weakly supervised manner. To this end, we propose a fully differentiable human deformation and rendering model, which allows us to compare the rendering of the human body model to the 2D image evidence and back-propagate the losses. For training, we first capture a video sequence in a calibrated multi-camera green screen studio (Sec. 3.1). Note that our multi-view video is only used during training. At test time we only require a single RGB video to perform dense non-rigid tracking.

#### 3.1. Template and Data Acquisition

**Character Model.** Our method relies on a person-specific 3D template model. We first scan the actor with a 3D scanner [79] to obtain the textured mesh. Then, it is automatically rigged to a kinematic skeleton, which is parameterized with joint angles \( \theta \in \mathbb{R}^{27} \), the camera relative rotation \( \alpha \in \mathbb{R}^3 \) and translation \( t \in \mathbb{R}^3 \). To model the non-rigid surface deformation, we automatically build an embedded deformation graph \( \mathcal{G} \) with \( K \) nodes following [77]. The nodes are parameterized with Euler angles \( \Lambda \in \mathbb{R}^{K \times 3} \) and translations \( T \in \mathbb{R}^{K \times 3} \). Similar to [32], we segment the mesh into different non-rigidity classes resulting in per-vertex rigidity weights \( s_i \). This allows us to model varying deformation behaviors of different surface materials, e.g., skin deforms less than clothing (see Eq. 13).

**Training Data.** To acquire the training data, we record a multi-view video of the actor doing various actions in a calibrated multi-camera studio with green screen. To provide weak supervision for the training, we first perform 2D pose detection on the sequences using OpenPose [19, 18, 72, 83] and apply temporal filtering. Then, we generate the foreground mask using color keying and compute the corresponding distance transform image \( D_{f,c} \) [12], where \( f \in [0, F] \) and \( c \in [0, C] \) denote the frame index and camera index, respectively. During training, we randomly sample one camera view \( c' \) and frame \( f' \) for which we crop the recorded image with a bounding box, based on the 2D joint detections. The final training input image \( I_{f',c'} \in \mathbb{R}_{256 \times 256 \times 3} \) is obtained by removing the background and augmenting the foreground with random brightness, hue, contrast and saturation changes. For simplicity, we describe the operation on frame \( f' \) and omit the subscript \( f' \) in following equations.

#### 3.2. Pose Network

In our *PoseNet*, we use ResNet50 [34] pretrained on ImageNet [25] as backbone and modify the last fully connected layer to output a vector containing the joint angles \( \theta \) and the camera relative root rotation \( \alpha \), given the input image \( I_{c'} \). Since generating the ground truth for \( \theta \) and \( \alpha \) is a non-trivial task, we propose weakly supervised training based on fitting the skeleton to multi-view 2D joint detections.

**Kinematics Layer.** To this end, we introduce a kinematics layer as the differentiable function that takes the joint angles \( \theta \) and the camera relative rotation \( \alpha \) and computes the positions \( P_{c'} \in \mathbb{R}^{M \times 3} \) of the \( M \) 3D landmarks attached to the skeleton (17 body joints and 4 face landmarks). Note that \( P_{c'} \) lives in a camera-root-relative coordinate system. In order to project the landmarks to other camera views, we
ne need to transform $P_{c',t}$ to the world coordinate system:

$$P_m = R_{c',t}^t P_{c',m} + t,$$

(1)

where $R_{c',t}$ is the rotation matrix of the input camera $c'$ and $t$ is the global translation of the skeleton.

**Global Alignment Layer.** To obtain the global translation $t$, we propose a global alignment layer that is attached to the kinematics layer. It localizes our skeleton model in the world space, such that the globally rotated landmarks $R_{c'}^t P_{c',m}$ project onto the corresponding detections in all camera views. This is done by minimizing the distance between the rotated landmarks $R_{c'}^t P_{c',m}$ and the corresponding rays cast from the camera origin $o_c$ to the 2D joint detections:

$$\sum_c \sum_m \sigma_{c,m} \| (R_{c'}^t P_{c',m} + t - o_c) \times d_{c,m} \|^2,$$

(2)

where $d_{c,m}$ is the direction of a ray from camera $c$ to the 2D joint detection $p_{c,m}$ corresponding to landmark $m$:

$$d_{c,m} = \frac{(E_c^{-1} p_{c,m})_{xyz} - o_c}{\|(E_c^{-1} p_{c,m})_{xyz} - o_c\|}.$$

(3)

Here, $E_c \in \mathbb{R}^{4 \times 4}$ is the projection matrix of camera $c$ and $p_{c,m} = (p_{c,m1}, 1, 1)^T$. Each point-to-line distance is weighted by the joint detection confidence $\sigma_{c,m}$, which is set to zero if below 0.4. The minimization problem of Eq. 2 can be solved in closed form:

$$t = W^{-1} \sum_{c,m} D_{c,m} (R_{c'}^t P_{c',m} - o_c) + o_c - R_{c'}^t P_{c',m},$$

(4)

where

$$W = \sum_c \sum_m I - D_{c,m}.$$

(5)

Here, $I$ is the $3 \times 3$ identity matrix and $D_{c,m} = d_{c,m} d_{c,m}^T$.

**Sparse Keypoint Loss.** Our 2D sparse keypoint loss for the PoseNet can be expressed as

$$L_{kp}(P) = \sum_c \sum_m \lambda_m \sigma_{c,m} \| \pi_c (P_m) - p_{c,m} \|^2,$$

(6)

which ensures that each landmark projects onto the corresponding 2D joint detections $p_{c,m}$ in all camera views. Here, $\pi_c$ is the projection function of camera $c$ and $\sigma_{c,m}$ is the same as in Eq. 2. $\lambda_m$ is a kinematic chain-based hierarchical weight which varies during training for better convergence (see the supplementary material for details).

**Pose Prior Loss.** To avoid unnatural poses, we impose a pose prior loss on the joint angles

$$L_{\text{prior}}(\theta) = \sum_{i=1}^{27} \Psi(\theta_i),$$

(7)

$$\Psi(x) = \begin{cases} (x - \theta_{\text{max},i})^2, & \text{if } x > \theta_{\text{max},i} \\ (\theta_{\text{min},i} - x)^2, & \text{if } x < \theta_{\text{min},i} \\ 0, & \text{otherwise} \end{cases},$$

(8)

that encourages that each joint angle $\theta_i$ stays in a range $[\theta_{\text{min},i}, \theta_{\text{max},i}]$ depending on the anatomic constraints.

### 3.3. Deformation Network

With the skeletal pose from PoseNet alone, the non-rigid deformation of the skin and clothes cannot be fully explained. Therefore, we disentangle the non-rigid deformation and the articulated skeletal motion. DefNet takes the
input image $I_c$ and regresses the non-rigid deformation parameterized with rotation angles $A$ and translation vectors $T$ of the nodes of the embedded deformation graph. $DefNet$ uses the same backbone architecture as $PoseNet$, while the last fully connected layer outputs a $6K$-dimensional vector reshaped to match the dimensions of $A$ and $T$. The weights of $PoseNet$ are fixed while training $DefNet$. Again, we do not use direct supervision on $A$ and $T$. Instead, we propose a deformation layer with differentiable rendering and use multi-view silhouette-based weak supervision.

**Deformation Layer.** The deformation layer takes $A$ and $T$ from $DefNet$ as input to non-rigidly deform the surface

$$Y_i = \sum_{k \in \mathcal{N}_{vn}(i)} w_{i,k}(R(A_k)(V_i - G_k) + G_k + T_k). \quad (9)$$

Here, $Y, V \in \mathbb{R}^{N \times 3}$ are the vertex positions of the deformed and undeformed template mesh, respectively. $w_{i,k}$ are vertex-to-node weights, but in contrast to $[77]$ we compute them based on geodesic distances. $G \in \mathbb{R}^{K \times 3}$ are the node positions of the undeformed graph, $\mathcal{N}_{vn}(i)$ is the set of nodes that influence vertex $i$, and $R(\cdot)$ is a function that converts the Euler angles to rotation matrices. We further apply the skeletal pose on the deformed mesh vertices to obtain the vertex positions in the input camera space

$$V_{c,i} = \sum_{k \in \mathcal{N}_{vn}(i)} w_{i,k}(R_{sk,k}(\theta, \alpha)Y_i + t_{sk,k}(\theta, \alpha)), \quad (10)$$

where the node rotation $R_{sk,k}$ and translation $t_{sk,k}$ are derived from the pose parameters using dual quaternion skinning $[44]$. Eq. $9$ and Eq. $10$ are differentiable with respect to pose and graph parameters. Thus, our layer can be integrated in the learning framework and gradients can be propagated to $DefNet$. So far, $V_{c,i}$ is still rotated relative to the camera $c'$ and located around the origin. To bring them to global space, we apply the inverse camera rotation and the global translation, defined in Eq. $4$, $V_i = R_{c'}^c V_{c,i} + t$.

**Non-rigid Silhouette Loss.** This loss encourages that the non-rigidly deformed mesh matches the multi-view silhouettes in all camera views. It can be formulated using the distance transform representation $[12]$

$$\mathcal{L}_{sil}(V) = \sum_{c} \sum_{i \in \mathcal{B}_c} \rho_{c,i} \| D_c (\pi_c (V_i)) \|^2. \quad (11)$$

Here, $\mathcal{B}_c$ is the set of vertices that lie on the boundary when the deformed 3D mesh is projected onto the distance transform image $D_c$ of camera $c$, and $\rho_{c,i}$ is a directional weighting $[32]$ that guides the gradient in $D_c$. The silhouette loss ensures that the boundary vertices project onto the zero-set of the distance transform, i.e., the foreground silhouette.

**Sparse Keypoint Graph Loss.** Only using the silhouette loss can lead to wrong mesh-to-image assignments, especially for highly articulated motions. To this end, we use a sparse keypoint loss to constrain the mesh deformation, which is similar to the keypoint loss for $PoseNet$ in Eq. $6$

$$\mathcal{L}_{kp}(M) = \sum_{i} \sum_{m} \sigma_{c,m} \| \pi_c (M_m) - p_{c,m} \|^2. \quad (12)$$

Differently from Eq. $6$, the deformed and posed landmarks $M$ are derived from the embedded deformation graph. To this end, we can deform and pose the canonical landmark positions by attaching them to its closest graph node $g$ in canonical pose with weight $w_{m,g} = 1.0$. Landmarks can then be deformed according to Eq. $9$, $10$, resulting in $M_g$ which is brought to global space via $M_m = R_{c'}^c M_{c',m} + t$.

**As-rigid-as-possible Loss.** To enforce local smoothness of the surface, we impose an as-rigid-as-possible loss $[75]$

$$\mathcal{L}_{arp}(A, T) = \sum_{k} \sum_{l \in \mathcal{N}_{c}(k)} u_{k,l} \| d_{k,l}(A, T) \|_1, \quad (13)$$

where $d_{k,l}(A, T) = R(A_k)(G_l - G_k) + T_k + G_k - (G_l + T_l)$.

$\mathcal{N}_{c}(k)$ is the set of indices of the neighbors of node $k$. In contrast to $[75]$, we propose weighting factors $u_{k,l}$ that influence the rigidity of respective parts of the graph. We derive $u_{k,l}$ by averaging all per-vertex rigidity weights $s_i$ $[32]$ for all vertices (see Sec. 3.1), which are connected to node $k$ or $l$. Thus, the mesh can deform either less or more depending on the surface material. For example, graph nodes that are mostly connected to vertices on a skirt can deform more freely than nodes that are mainly connected to vertices on the skin.

**3.4. In-the-wild Domain Adaptation**

Since our training set is captured in a green screen studio and our test set is captured in the wild, there is a significant domain gap between them, due to different lighting conditions and camera response functions. To improve the performance of our method on in-the-wild images, we fine-tune our networks on the monocular test images for a small number of iterations using the same 2D keypoint and silhouette losses as before, but only on a single view. This drastically improves the performance at test time as shown in the supplemental material.

**4. Results**

All our experiments were performed on a machine with an NVIDIA Tesla V100 GPU. A forward pass of our method takes around 50ms, which breaks down to 25ms for $PoseNet$ and 25ms for $DefNet$. During testing, we use the off-the-shelf video segmentation method of $[16]$ to remove the background in the input image. Our method requires OpenPose’s 2D joint detections $[19, 18, 72, 83]$ as
input during testing to crop the frames and to obtain the 3D global translation with our global alignment layer. Finally, we temporally smooth the output mesh vertices with a Gaussian kernel of size 5 frames.

**Dataset.** We evaluate our approach on 4 subjects (S1 to S4) with varying types of apparel. For qualitative evaluation, we recorded 13 in-the-wild sequences in different indoor and outdoor environments shown in Fig. 3. For quantitative evaluation, we captured 4 sequences in a calibrated multi-camera green screen studio (see Fig. 4), for which we computed the ground truth 3D joint locations using the multi-view motion capture software, The Captury [20], and we use a color keying algorithm for ground truth foreground segmentation. All sequences contain a large variety of motions, ranging from simple ones like walking up to more difficult ones like fast dancing or baseball pitching. We will release the dataset for future research.

**Qualitative Comparisons.** Fig. 3 shows our qualitative results on in-the-wild test sequences with various clothing styles, poses and environments. Our reconstructions not only precisely overlay with the input images, but also look plausible from arbitrary 3D view points. In Fig. 5, we qualitatively compare our approach to the related human capture and reconstruction methods [42, 32, 70, 99]. In terms of the shape representation, our method is most

Figure 3. Qualitative results. Each row shows results for a different person with varying types of apparel. We visualize input frames and our reconstruction overlayed to the corresponding frame. Note that our results precisely overlay to the input. Further, we show our reconstructions from a virtual 3D viewpoint. Note that they also look plausible in 3D.

Figure 4. Results on our evaluation sequences where input views (IV) and reference views (RV) are available. Note that our reconstruction also precisely overlays on RV even though they are not used for tracking.
Table 1. Skeletal pose accuracy. Note that we are consistently better than other monocular approaches. Moreover, we are even close to the multi-view baseline.

<table>
<thead>
<tr>
<th>Method</th>
<th>GLE↑ (in mm)</th>
<th>3DPCK↑ (in %)</th>
<th>AUC↑ (in %)</th>
<th>MPJPE↓ (in mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VNect [53]</td>
<td>-</td>
<td>66.06</td>
<td>28.02</td>
<td>77.19</td>
</tr>
<tr>
<td>HMR [42]</td>
<td>-</td>
<td>82.39</td>
<td>43.61</td>
<td>72.61</td>
</tr>
<tr>
<td>HMMR [43]</td>
<td>-</td>
<td>87.48</td>
<td>45.33</td>
<td>72.40</td>
</tr>
<tr>
<td>LiveCap [32]</td>
<td>317.03</td>
<td>71.13</td>
<td>37.90</td>
<td>92.84</td>
</tr>
<tr>
<td>Ours</td>
<td>91.08</td>
<td>98.43</td>
<td>58.71</td>
<td>49.11</td>
</tr>
<tr>
<td>MVBL</td>
<td>76.03</td>
<td>99.17</td>
<td>57.79</td>
<td>45.44</td>
</tr>
</tbody>
</table>

For every 10th frame: 1) We evaluate the root joint position error or global localization error \( (GLE) \) to measure how good the skeleton is placed in global 3D space. Note that \( GLE \) can only be evaluated for LiveCap [32] and ours, since other methods only produce up-to-scale depth. 2) To evaluate the accuracy of the pose estimation, we report the 3D percentage of correct keypoints \( (3DPCK) \) with a threshold of 150\( \text{mm} \) of the root aligned poses and the area under the 3DPCK curve \( (AUC) \). 3) To factor out the errors in the global rotation, we also report the mean per joint position error \( (MPJPE) \) after Procrustes alignment. We compare our approach against the state-of-the-art pose estimation approaches including VNect [53], HMR [42], HMMR [43], and LiveCap [32]. We also compare to a multi-view baseline approach \( (MVBL) \), where we use our differentiable skeleton model in an optimization framework to solve for the pose per frame using the proposed multi-view losses. We see from Tab. 3 that our approach outperforms the related monocular methods in all metrics by a large margin and is even close to \( MVBL \) although our method only takes a single input as input. We further compare to VNect [53] fine-tuned on our training images for \( S1 \). To this end, we compute the 3D joint position using The Captury [20] to provide ground truth supervision for VNect. On the evaluation sequence for \( S1 \), the fine-tuned VNect achieved 95.66% 3DPCK, 52.13% AUC and 47.16\( \text{mm} \) MPJPE. This shows our weakly supervised approach yields comparable or better results than supervised methods in the person-specific setting. However, our approach does not require 3D ground truth annotation that is difficult to obtain, even for only sparse keypoints, let alone the dense surfaces.

Surface Reconstruction Accuracy. To evaluate the accuracy of the regressed non-rigid deformations, we compute the intersection over union \( (IoU) \) between the ground truth foreground masks and the 2D projection of the estimated foreground masks.
shape on S1 and S4 for every 100th frame. We evaluate the IoU on all views, on all views expect the input view, and on the input view which we refer to as AMVIoU, RVIoU and SVIoU, respectively. To factor out the errors in global localization, we apply the ground truth translation to the reconstructed geometries. For DeepHuman [99] and PiFu [70], we cannot report the AMVIoU and RVIoU, since we cannot overlay their results on reference views as discussed before. Further, PiFu [70] by design achieves perfect overlay on the input view, since they regress the depth for each foreground pixel. However, their reconstruction does not reflect the true 3D geometry (see Fig. 5). Therefore, it is meaningless to report their SVIoU. Similarly, DeepHuman [99] achieves high SVIoU, due to their volumetric representation. But their results are often wrong, when looking from side views. In contrast, our method consistently outperforms all other approaches in terms of AMVIoU and RVIoU, which shows the high accuracy of our method in recovering the 3D geometry. Further, we are again close to the multi-view baseline.

Ablation Study. To evaluate the importance of the number of cameras, the number of training images, and our DefNet, we performed an ablation study on S4 in Tab. 3. 1) In the first group of Tab. 3, we train our networks with supervision using 1 to 7 views. We can see that adding more views consistently improves the quality of the estimated poses and deformations. The most significant improvement is from one to two cameras. This is not surprising, since the single camera settings is inherently ambiguous. 2) In the second group of Tab. 3, we reduce the training data to 1/2 and 1/4. We can see that the more frames with different poses and deformations are seen during training, the better the reconstruction quality is. This is expected since a larger number of frames may better sample the possible space of poses and deformations. 3) In the third group of Tab. 3, we evaluate the AMVIoU on the template mesh animated with the results of PoseNet, which refer to as PoseNet-only. One can see that on average, the AMVIoU is improved by around 4%. Since most non-rigid deformations rather happen locally, the difference is visually even more significant as shown in Fig. 6. Especially, the skirt is correctly deformed according to the input image whereas the PoseNet-only result cannot fit the input due to the limitation of skinning.

5. Conclusion

We have presented a learning-based approach for monocular dense human performance capture using only weak multi-view supervision. In contrast to existing methods, our approach directly regresses poses and surface deformations from neural networks, produces temporal surface correspondences, preserves the skeletal structure of the human body, and can handle loose clothing. Our qualitative and quantitative results in different scenarios show that our method produces more accurate 3D reconstruction of pose and non-rigid deformation than existing methods. In the future, we plan to incorporate hands and the face to our mesh representation to enable joint tracking of body, facial expressions and hand gestures. We are also interested in physically more correct multi-layered representations to model the garments even more realistically.

Acknowledgements. This work was funded by the ERC Consolidator Grant 4DRepLy (770784) and the Deutsche Forschungsgemeinschaft (Project Nr. 409792180, Emmy Noether Programme, project: Real Virtual Humans).

<table>
<thead>
<tr>
<th>Method</th>
<th>AMVIoU</th>
<th>RVIoU</th>
<th>SVIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMR [42]</td>
<td>62.25</td>
<td>61.7</td>
<td>68.85</td>
</tr>
<tr>
<td>HMMR [43]</td>
<td>65.98</td>
<td>65.58</td>
<td>70.77</td>
</tr>
<tr>
<td>LiveCap [32]</td>
<td>56.02</td>
<td>54.21</td>
<td>77.75</td>
</tr>
<tr>
<td>DeepHuman [99]</td>
<td>89.26</td>
<td>91.57</td>
<td>89.26</td>
</tr>
<tr>
<td>Ours</td>
<td>91.74</td>
<td>91.72</td>
<td>92.02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>AMVIoU</th>
<th>RVIoU</th>
<th>SVIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMR [42]</td>
<td>85.1</td>
<td>66.66</td>
<td>70.84</td>
</tr>
<tr>
<td>HMMR [43]</td>
<td>63.79</td>
<td>63.29</td>
<td>70.23</td>
</tr>
<tr>
<td>LiveCap [32]</td>
<td>59.96</td>
<td>59.02</td>
<td>71.56</td>
</tr>
<tr>
<td>DeepHuman [99]</td>
<td>84.15</td>
<td>84.15</td>
<td>84.15</td>
</tr>
<tr>
<td>Ours</td>
<td>82.53</td>
<td>82.22</td>
<td>86.66</td>
</tr>
<tr>
<td>MVBL</td>
<td>88.14</td>
<td>88.03</td>
<td>89.06</td>
</tr>
</tbody>
</table>

Table 2. Surface deformation accuracy. We outperform all other monocular methods and are even close to the multi-view baseline.

<table>
<thead>
<tr>
<th>Method</th>
<th>3DPCK</th>
<th>AMVIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 camera view</td>
<td>62.11</td>
<td>65.11</td>
</tr>
<tr>
<td>2 camera views</td>
<td>93.52</td>
<td>78.44</td>
</tr>
<tr>
<td>3 camera views</td>
<td>94.70</td>
<td>79.75</td>
</tr>
<tr>
<td>7 camera views</td>
<td>95.95</td>
<td>81.73</td>
</tr>
<tr>
<td>6500 frames</td>
<td>85.19</td>
<td>73.41</td>
</tr>
<tr>
<td>13000 frames</td>
<td>92.25</td>
<td>78.97</td>
</tr>
<tr>
<td>PoseNet-only</td>
<td>96.74</td>
<td>78.51</td>
</tr>
<tr>
<td>Ours(14 views, 26000 frames)</td>
<td>96.74</td>
<td>82.53</td>
</tr>
</tbody>
</table>

Table 3. Ablation study. We evaluate the number of cameras and the number of frames used during training in terms of the 3DPCK and AMVIoU metrics. Adding more cameras and frames consistently improves the quality of reconstruction. Further, DefNet improves the AMVIoU compared to pure pose estimation.

Figure 6. PoseNet + DefNet vs. PoseNet-only. DefNet can deform the template to accurately match the input, especially for loose clothing. In addition, DefNet also corrects slight errors in the pose and typical skinning artifacts.
References


