

DeepCap: Monocular Human Performance Capture Using Weak Supervision

Marc Habermann^{1,2} Weipeng Xu^{1,2} Michael Zollhoefer³ Gerard Pons-Moll^{1,2} Christian Theobalt^{1,2}

¹Max Planck Institute for Informatics, ²Saarland Informatics Campus, ³Stanford 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 performance 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].

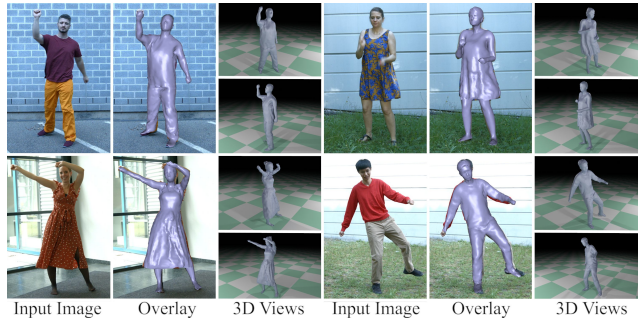


Figure 1. We present the first learning-based approach for dense monocular human performance capture using weak multi-view supervision that not only predicts the pose but also the space-time coherent non-rigid deformations of the model surface.

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 key-point 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

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]). Instead, our method is based on person-specific templates including clothes and employs deep learning to predict clothing deformation based on monocular video directly.

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. DeepHuman [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 $\mathbf{t} \in \mathbb{R}^3$. 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 $\mathbf{A} \in \mathbb{R}^{K \times 3}$ and translations $\mathbf{T} \in \mathbb{R}^{K \times 3}$. 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 θ and the camera relative root rotation α , given the input image $I_{c'}$. Since generating the ground truth for θ and α 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 θ and the camera relative rotation α and computes the positions $\mathbf{P}_{c'} \in \mathbb{R}^{M \times 3}$ of the M 3D landmarks attached to the skeleton (17 body joints and 4 face landmarks). Note that $\mathbf{P}_{c'}$ lives in a camera-root-relative coordinate system. In order to project the landmarks to other camera views, we

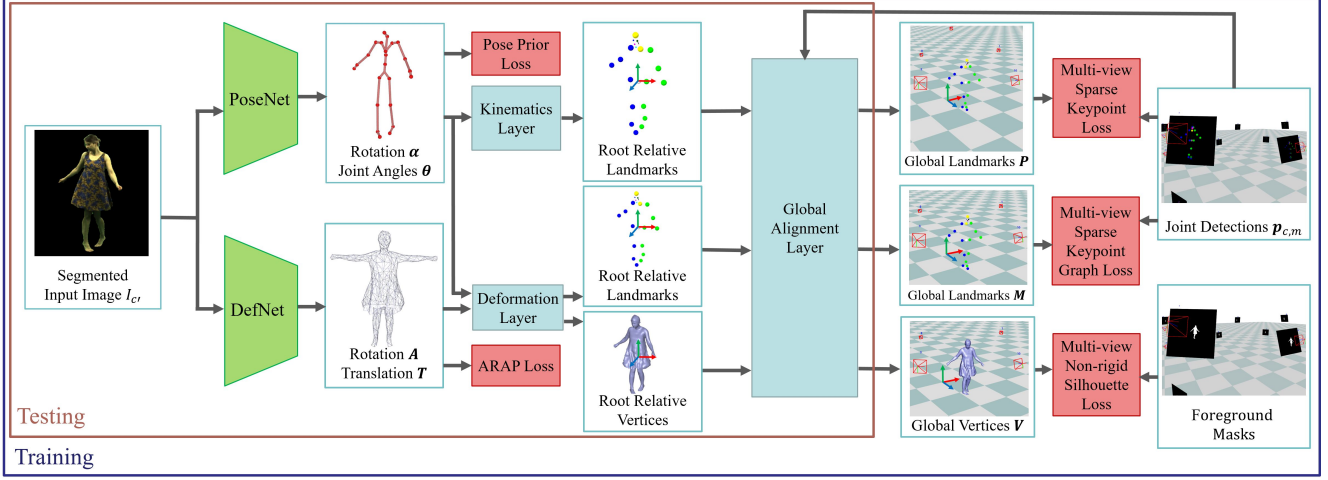


Figure 2. Overview of our approach. Our method takes a single segmented image as input. First, our pose network, *PoseNet*, is trained to predict the joint angles and the camera relative rotation using sparse multi-view 2D joint detections as weak supervision. Second, the deformation network, *DefNet*, is trained to regress embedded graph rotation and translation parameters to account for non-rigid deformations. To train *DefNet*, multi-view 2D joint detections and silhouettes are used for supervision.

need to transform $\mathbf{P}_{c'}$ to the world coordinate system:

$$\mathbf{P}_m = \mathbf{R}_{c'}^T \mathbf{P}_{c',m} + \mathbf{t}, \quad (1)$$

where $\mathbf{R}_{c'}$ is the rotation matrix of the input camera c' and \mathbf{t} is the global translation of the skeleton.

Global Alignment Layer. To obtain the global translation \mathbf{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 $\mathbf{R}_{c'}^T \mathbf{P}_{c',m}$ project onto the corresponding detections in all camera views. This is done by minimizing the distance between the rotated landmarks $\mathbf{R}_{c'}^T \mathbf{P}_{c',m}$ and the corresponding rays cast from the camera origin \mathbf{o}_c to the 2D joint detections:

$$\sum_c \sum_m \sigma_{c,m} \|(\mathbf{R}_{c'}^T \mathbf{P}_{c',m} + \mathbf{t} - \mathbf{o}_c) \times \mathbf{d}_{c,m}\|^2, \quad (2)$$

where $\mathbf{d}_{c,m}$ is the direction of a ray from camera c to the 2D joint detection $\mathbf{p}_{c,m}$ corresponding to landmark m :

$$\mathbf{d}_{c,m} = \frac{(\mathbf{E}_c^{-1} \tilde{\mathbf{p}}_{c,m})_{xyz} - \mathbf{o}_c}{\|(\mathbf{E}_c^{-1} \tilde{\mathbf{p}}_{c,m})_{xyz} - \mathbf{o}_c\|}. \quad (3)$$

Here, $\mathbf{E}_c \in \mathbb{R}^{4 \times 4}$ is the projection matrix of camera c and $\tilde{\mathbf{p}}_{c,m} = (\mathbf{p}_{c,m}, 1, 1)^T$. 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:

$$\mathbf{t} = \mathbf{W}^{-1} \sum_{c,m} \mathbf{D}_{c,m} (\mathbf{R}_{c'}^T \mathbf{P}_{c',m} - \mathbf{o}_c) + \mathbf{o}_c - \mathbf{R}_{c'}^T \mathbf{P}_{c',m}, \quad (4)$$

where

$$\mathbf{W} = \sum_c \sum_m \mathbf{I} - \mathbf{D}_{c,m}. \quad (5)$$

Here, \mathbf{I} is the 3×3 identity matrix and $\mathbf{D}_{c,m} = \mathbf{d}_{c,m} \mathbf{d}_{c,m}^T$. Note that the operation in Eq. 4 is differentiable with respect to the landmark position $\mathbf{P}_{c'}$.

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

$$\mathcal{L}_{kp}(\mathbf{P}) = \sum_c \sum_m \lambda_m \sigma_{c,m} \|\pi_c(\mathbf{P}_m) - \mathbf{p}_{c,m}\|^2, \quad (6)$$

which ensures that each landmark projects onto the corresponding 2D joint detections $\mathbf{p}_{c,m}$ in all camera views. Here, π_c is the projection function of camera c and $\sigma_{c,m}$ is the same as in Eq. 2. λ_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

$$\mathcal{L}_{limit}(\boldsymbol{\theta}) = \sum_{i=1}^{27} \Psi(\theta_i) \quad (7)$$

$$\Psi(x) = \begin{cases} (x - \theta_{\max,i})^2, & \text{if } x > \theta_{\max,i} \\ (\theta_{\min,i} - x)^2, & \text{if } x < \theta_{\min,i} \\ 0, & \text{otherwise} \end{cases}, \quad (8)$$

that encourages that each joint angle θ_i stays in a range $[\theta_{\min,i}, \theta_{\max,i}]$ 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 \mathbf{A} and translation vectors \mathbf{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 \mathbf{A} and \mathbf{T} . The weights of *PoseNet* are fixed while training *DefNet*. Again, we do not use direct supervision on \mathbf{A} and \mathbf{T} . Instead, we propose a deformation layer with differentiable rendering and use multi-view silhouette-based weak supervision.

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

$$\mathbf{Y}_i = \sum_{k \in \mathcal{N}_{\text{vn}}(i)} w_{i,k} (R(\mathbf{A}_k)(\hat{\mathbf{V}}_i - \mathbf{G}_k) + \mathbf{G}_k + \mathbf{T}_k). \quad (9)$$

Here, $\mathbf{Y}, \hat{\mathbf{V}} \in \mathbb{R}^{N \times 3}$ 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. $\mathbf{G} \in \mathbb{R}^{K \times 3}$ are the node positions of the undeformed graph, $\mathcal{N}_{\text{vn}}(i)$ 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

$$\mathbf{V}_{c',i} = \sum_{k \in \mathcal{N}_{\text{vn}}(i)} w_{i,k} (R_{\text{sk},k}(\boldsymbol{\theta}, \boldsymbol{\alpha}) \mathbf{Y}_i + t_{\text{sk},k}(\boldsymbol{\theta}, \boldsymbol{\alpha})), \quad (10)$$

where the node rotation $R_{\text{sk},k}$ and translation $t_{\text{sk},k}$ 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, $\mathbf{V}_{c',i}$ 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, $\mathbf{V}_i = \mathbf{R}_{c'}^T \mathbf{V}_{c',i} + \mathbf{t}$.

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]

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

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

$$\mathcal{L}_{\text{kp}}(\mathbf{M}) = \sum_c \sum_m \sigma_{c,m} \|\pi_c(\mathbf{M}_m) - \mathbf{p}_{c,m}\|^2. \quad (12)$$

Differently from Eq. 6, the deformed and posed landmarks \mathbf{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 $\mathbf{M}_{c'}$ which is brought to global space via $\mathbf{M}_m = \mathbf{R}_{c'}^T \mathbf{M}_{c',m} + \mathbf{t}$. **As-rigid-as-possible Loss.** To enforce local smoothness of the surface, we impose an as-rigid-as-possible loss [75]

$$\mathcal{L}_{\text{arap}}(\mathbf{A}, \mathbf{T}) = \sum_k \sum_{l \in \mathcal{N}_n(k)} u_{k,l} \|d_{k,l}(\mathbf{A}, \mathbf{T})\|_1, \quad (13)$$

where

$$d_{k,l}(\mathbf{A}, \mathbf{T}) = R(\mathbf{A}_k)(\mathbf{G}_l - \mathbf{G}_k) + \mathbf{T}_k + \mathbf{G}_k - (\mathbf{G}_l + \mathbf{T}_l).$$

$\mathcal{N}_n(k)$ 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



Figure 3. Qualitative results. Each row shows results for a different person with varying types of apparel. We visualize input frames and our reconstruction overlaid 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.

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

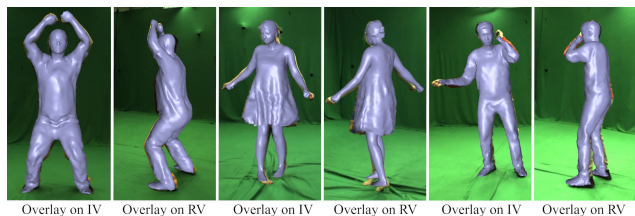


Figure 4. Results on our evaluation sequences where input views (IV) and reference views (RV) are available. Note that our reconstruction also precisely overlays on RV even though they are not used for tracking.

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

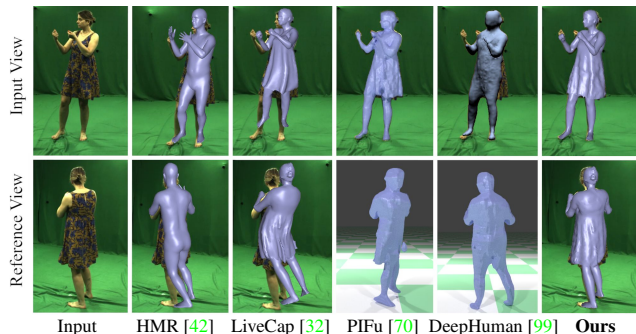


Figure 5. Qualitative comparison to other methods [42, 32, 70, 99]. Note that our results overlay more accurately to the input view and also look more plausible from a reference view that was not used for tracking. Ground truth global translation is used to match the reference view for the results of [42, 32]. Since PIFu [70] and DeepHuman [99] output meshes with varying topology in a canonical volume without an attached root, it is not possible to apply the ground truth translation and therefore we show the reference view without overlay.

closely related to LiveCap [32] that also uses a person-specific template. Since they non-rigidly fit the template only to the monocular input view, their results do not faithfully depict the deformation in other view points. Further, their pose estimation severely suffers from the monocular ambiguities, whereas our pose results are more robust and accurate (see supplemental video). Comparing to the other three methods [42, 70, 99] that are trained for general subjects, our approach has the following advantages: First, our method recovers the non-rigid deformations of humans in general clothes whereas the parametric model-based approaches [42, 43] only recover naked body shape. Second, our method directly provides surface correspondences over time which is important for AR/VR applications (see supplemental video). In contrast, the results of implicit representation-based methods, PIFu [70] and DeepHuman [99], lack temporal surface correspondences and do not preserve the skeletal structure of the human body, *i.e.*, they often exhibit missing arms and disconnected geometry. Furthermore, DeepHuman [99] only recovers a coarse shape in combination with a normal image of the input view, while our method can recover medium-level detailed geometry that looks plausible from all views. Last but not least, all these existing methods have problems when overlaying their reconstructions on the reference view, even though some of the methods show a very good overlay on the input view. In contrast, our approach reconstructs accurate 3D geometry, and therefore, our results can precisely overlay on the reference views (also see Fig. 4).

Skeletal Pose Accuracy. We quantitatively compare our pose results (output of *PoseNet*) to existing pose estimation methods on *S1* and *S4*. To account for different types of apparel, we choose *S1* wearing trousers and a T-shirt and *S4* wearing a short dress. We rescale the bone length for

| <i>MPJPE/GLE (in mm) and 3DPCK/AUC (in %) on S1</i> | | | | |
|---|--------------|--------------|--------------|--------------|
| Method | GLE↓ | 3DPCK↑ | AUC↑ | MPJPE↓ |
| VNect [53] | - | 66.06 | 28.02 | 77.19 |
| HMR [42] | - | 82.39 | 43.61 | 72.61 |
| HMMR [43] | - | 87.48 | 45.33 | 72.40 |
| LiveCap [32] | 317.01 | 71.13 | 37.90 | 92.84 |
| Ours | 91.08 | 98.43 | 58.71 | 49.11 |
| MVBL | 76.03 | 99.17 | 57.79 | 45.44 |

| <i>MPJPE/GLE (in mm) and 3DPCK/AUC (in %) on S4</i> | | | | |
|---|--------------|--------------|--------------|--------------|
| Method | GLE↓ | 3DPCK↑ | AUC↑ | MPJPE↓ |
| VNect [53] | - | 82.06 | 42.73 | 72.62 |
| HMR [42] | - | 86.88 | 43.91 | 73.63 |
| HMMR [43] | - | 82.80 | 41.18 | 77.41 |
| LiveCap [32] | 248.67 | 75.11 | 37.35 | 83.48 |
| Ours | 96.56 | 96.74 | 59.25 | 45.40 |
| MVBL | 75.82 | 96.20 | 57.27 | 45.12 |

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.

all methods to the ground truth and evaluate the following metrics on the 14 commonly used joints [53] 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 150mm 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 can 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 image 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 Capture [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 47.16mm 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

| AMVloU, RVloU, and SVloU (in %) on S1 sequence | | | |
|--|-------------------|------------------|------------------|
| Method | AMVloU \uparrow | RVloU \uparrow | SVloU \uparrow |
| HMR [42] | 62.25 | 61.7 | 68.85 |
| HMMR [43] | 65.98 | 65.58 | 70.77 |
| LiveCap [32] | 56.02 | 54.21 | 77.75 |
| DeepHuman [99] | - | - | 91.57 |
| Ours | 87.2 | 87.03 | 89.26 |
| MVBL | 91.74 | 91.72 | 92.02 |

| AMVloU, RVloU, and SVloU (in %) on S4 sequence | | | |
|--|-------------------|------------------|------------------|
| Method | AMVloU \uparrow | RVloU \uparrow | SVloU \uparrow |
| HMR [42] | 65.1 | 64.66 | 70.84 |
| HMMR [43] | 63.79 | 63.29 | 70.23 |
| LiveCap [32] | 59.96 | 59.02 | 72.16 |
| DeepHuman [99] | - | - | 84.15 |
| Ours | 82.53 | 82.22 | 86.66 |
| MVBL | 88.14 | 88.03 | 89.66 |

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

shape on *S1* and *S4* for every 100th frame. We evaluate the IoU on *all views*, on *all views except the input view*, and on the *input view* which we refer to as *AMVloU*, *RVloU* and *SVloU*, 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 *AMVloU* and *RVloU*, 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 *SVloU*. Similarly, DeepHuman [99] achieves high *SVloU*, 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 *AMVloU* and *RVloU*, 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 *AMVloU* on the template mesh animated with the results of *PoseNet*, which we refer to as *PoseNet-only*. One can see that on average, the *AMVloU* is improved by around 4%. Since most non-rigid deformations rather happen locally,

| 3DPCK and AMVloU (in %) on S4 sequence | | |
|--|------------------|-------------------|
| Method | 3DPCK \uparrow | AMVloU \uparrow |
| 1 camera view | 62.11 | 65.11 |
| 2 camera views | 93.52 | 78.44 |
| 3 camera views | 94.70 | 79.75 |
| 7 camera views | 95.95 | 81.73 |
| 6500 frames | 85.19 | 73.41 |
| 13000 frames | 92.25 | 78.97 |
| PoseNet-only | 96.74 | 78.51 |
| Ours(14 views, 26000 frames) | 96.74 | 82.53 |

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

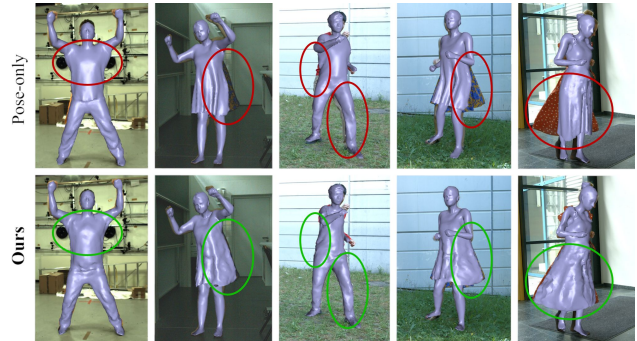


Figure 6. *PoseNet + DefNet* vs. *PoseNet-only*. *DefNet* can deform the template to accurately match the input, especially for loose clothing. In addition, *DefNet* also corrects slight errors in the pose and typical skinning artifacts.

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 clothes. 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).

References

- [1] B. Allain, J.-S. Franco, and E. Boyer. An Efficient Volumetric Framework for Shape Tracking. In *CVPR 2015 - IEEE International Conference on Computer Vision and Pattern Recognition*, pages 268–276, Boston, United States, June 2015. IEEE. 2
- [2] T. Alldieck, M. Magnor, B. L. Bhatnagar, C. Theobalt, and G. Pons-Moll. Learning to reconstruct people in clothing from a single RGB camera. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1175–1186, Jun 2019. 1, 2
- [3] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3d people models. In *IEEE Conference on Computer Vision and Pattern Recognition*. CVPR Spotlight Paper. 1, 2
- [4] T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Detailed human avatars from monocular video. In *International Conference on 3D Vision*, pages 98–109, Sep 2018. 2
- [5] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor. Tex2shape: Detailed full human body geometry from a single image. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2019. 1
- [6] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis. SCAPE: Shape Completion and Animation of People. *ACM Transactions on Graphics*, 24(3):408–416, 2005. 2
- [7] A. O. Bălan and M. J. Black. The naked truth: Estimating body shape under clothing. In *European Conference on Computer Vision*, pages 15–29. Springer, 2008. 2
- [8] A. O. Balan, L. Sigal, M. J. Black, J. E. Davis, and H. W. Haussecker. Detailed human shape and pose from images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2007. 2
- [9] B. L. Bhatnagar, G. Tiwari, C. Theobalt, and G. Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2019. 1, 2
- [10] F. Bogo, M. J. Black, M. Loper, and J. Romero. Detailed full-body reconstructions of moving people from monocular RGB-D sequences. In *International Conference on Computer Vision (ICCV)*, pages 2300–2308, Dec. 2015. 2
- [11] F. Bogo, A. Kanazawa, C. Lassner, P. Gehler, J. Romero, and M. J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European Conference on Computer Vision (ECCV)*, 2016. 2
- [12] G. Borgefors. Distance transformations in digital images. *Computer Vision, Graphics, and Image Processing*, 34(3):344 – 371, 1986. 3, 5
- [13] M. Bray, P. Kohli, and P. H. Torr. Posecut: Simultaneous segmentation and 3d pose estimation of humans using dynamic graph-cuts. In *European conference on computer vision*, pages 642–655. Springer, 2006. 1, 3
- [14] T. Brox, B. Rosenhahn, D. Cremers, and H.-P. Seidel. High accuracy optical flow serves 3-d pose tracking: exploiting contour and flow based constraints. In *European Conference on Computer Vision*, pages 98–111. Springer, 2006. 1
- [15] T. Brox, B. Rosenhahn, J. Gall, and D. Cremers. Combined region and motion-based 3d tracking of rigid and articulated objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3):402–415, 2010. 1, 3
- [16] S. Caelles, K.-K. Maninis, J. Pont-Tuset, L. Leal-Taixé, D. Cremers, and L. Van Gool. One-shot video object segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 5
- [17] C. Cagniart, E. Boyer, and S. Ilic. Free-form mesh tracking: a patch-based approach. In *Computer Vision and Pattern Recognition (CVPR)*, 2010 IEEE Conference on, pages 1339–1346. IEEE, 2010. 1, 3
- [18] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*, 2018. 3, 5
- [19] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 3, 5
- [20] The Capture. <http://www.thecapture.com/>. 6, 7
- [21] J. Carranza, C. Theobalt, M. A. Magnor, and H.-P. Seidel. Free-viewpoint video of human actors. *ACM Trans. Graph.*, 22(3), July 2003. 3
- [22] J. Chibane, T. Alldieck, and G. Pons-Moll. Implicit functions in feature space for 3d shape reconstruction and completion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020. 1
- [23] A. Collet, M. Chuang, P. Sweeney, D. Gillett, D. Evseev, D. Calabrese, H. Hoppe, A. Kirk, and S. Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (TOG)*, 34(4):69, 2015. 2
- [24] E. De Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun. Performance capture from sparse multi-view video. In *ACM Transactions on Graphics (TOG)*, volume 27, page 98. ACM, 2008. 1, 3
- [25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 3
- [26] M. Dou, P. Davidson, S. R. Fanello, S. Khamis, A. Kowdle, C. Rhemann, V. Tankovich, and S. Izadi. Motion2fusion: Real-time volumetric performance capture. *ACM Trans. Graph.*, 36(6):246:1–246:16, Nov. 2017. 2
- [27] M. Dou, S. Khamis, Y. Degtyarev, P. Davidson, S. R. Fanello, A. Kowdle, S. O. Escolano, C. Rhemann, D. Kim, J. Taylor, et al. Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (TOG)*, 35(4):114, 2016. 2
- [28] V. Gabeur, J.-S. Franco, X. Martin, C. Schmid, and G. Rogez. Moulding humans: Non-parametric 3d human shape estimation from single images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2232–2241, 2019. 1
- [29] J. Gall, C. Stoll, E. De Aguiar, C. Theobalt, B. Rosenhahn, and H.-P. Seidel. Motion capture using joint skeleton tracking and surface estimation. In *Computer Vision and Pat-*

- tern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1746–1753. IEEE, 2009. 1, 3
- [30] K. Guo, J. Taylor, S. Fanello, A. Tagliasacchi, M. Dou, P. Davidson, A. Kowdle, and S. Izadi. Twinfusion: High framerate non-rigid fusion through fast correspondence tracking. 09 2018. 2
- [31] K. Guo, F. Xu, T. Yu, X. Liu, Q. Dai, and Y. Liu. Real-time geometry, albedo, and motion reconstruction using a single rgb-d camera. *ACM Transactions on Graphics (TOG)*, 36(3):32, 2017. 2
- [32] M. Habermann, W. Xu, M. Zollhoefer, G. Pons-Moll, and C. Theobalt. Livecap: Real-time human performance capture from monocular video. *ACM Trans. Graph.*, 2019. 1, 3, 5, 6, 7, 8
- [33] N. Hasler, H. Ackermann, B. Rosenhahn, T. Thormählen, and H.-P. Seidel. Multilinear pose and body shape estimation of dressed subjects from image sets. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1823–1830. IEEE, 2010. 2
- [34] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016*. 3
- [35] T. Helten, M. Muller, H.-P. Seidel, and C. Theobalt. Real-time body tracking with one depth camera and inertial sensors. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2013. 2
- [36] C.-H. Huang, B. Allain, J.-S. Franco, N. Navab, S. Ilic, and E. Boyer. Volumetric 3d tracking by detection. In *Proc. CVPR, 2016*. 2
- [37] Y. Huang, F. Bogo, C. Lassner, A. Kanazawa, P. V. Gehler, J. Romero, I. Akhter, and M. J. Black. Towards accurate marker-less human shape and pose estimation over time. In *International Conference on 3D Vision (3DV)*, 2017. 2
- [38] Z. Huang, T. Li, W. Chen, Y. Zhao, J. Xing, C. LeGendre, L. Luo, C. Ma, and H. Li. Deep Volumetric Video From Very Sparse Multi-View Performance Capture. In *Proceedings of the 15th European Conference on Computer Vision*. Computer Vision Foundation, Sept. 2018. 2
- [39] M. Innmann, M. Zollhöfer, M. Nießner, C. Theobalt, and M. Stamminger. VolumeDeform: Real-time Volumetric Non-rigid Reconstruction. October 2016. 2
- [40] H. Joo, T. Simon, and Y. Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. *CoRR*, abs/1801.01615, 2018. 2
- [41] P. Kadlecik, A.-E. Ichim, T. Liu, J. Krivanek, and L. Kavan. Reconstructing personalized anatomical models for physics-based body animation. *ACM Trans. Graph.*, 35(6), 2016. 2
- [42] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. End-to-end recovery of human shape and pose. *CoRR*, abs/1712.06584, 2017. 2, 6, 7, 8
- [43] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik. Learning 3d human dynamics from video. In *Computer Vision and Pattern Recognition (CVPR), 2019*. 2, 7, 8
- [44] L. Kavan, S. Collins, J. Žára, and C. O’Sullivan. Skinning with dual quaternions. In *Proceedings of the 2007 Symposium on Interactive 3D Graphics and Games, I3D ’07, 2007*. 5
- [45] M. Kim, G. Pons-Moll, S. Pujades, S. Bang, J. Kim, M. Black, and S.-H. Lee. Data-driven physics for human soft tissue animation. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 36(4), 2017. 2
- [46] N. Kolotouros, G. Pavlakos, and K. Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR, 2019*. 2
- [47] A. Kowdle, C. Rhemann, S. Fanello, A. Tagliasacchi, J. Taylor, P. Davidson, M. Dou, K. Guo, C. Keskin, S. Khamis, D. Kim, D. Tang, V. Tankovich, J. Valentin, and S. Izadi. The need 4 speed in real-time dense visual tracking. In *SIGGRAPH Asia 2018 Technical Papers, SIGGRAPH Asia ’18*, pages 220:1–220:14, New York, NY, USA, 2018. ACM. 2
- [48] C. Lassner, J. Romero, M. Kiefel, F. Bogo, M. J. Black, and P. V. Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *Proc. CVPR, 2017*. 2
- [49] V. Leroy, J.-S. Franco, and E. Boyer. Multi-View Dynamic Shape Refinement Using Local Temporal Integration. In *IEEE, International Conference on Computer Vision 2017, Venice, Italy, Oct. 2017*. 2
- [50] Y. Liu, C. Stoll, J. Gall, H.-P. Seidel, and C. Theobalt. Markerless motion capture of interacting characters using multi-view image segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1249–1256. IEEE, 2011. 1, 3
- [51] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 2
- [52] Q. Ma, J. Yang, A. Ranjan, S. Pujades, G. Pons-Moll, S. Tang, and M. Black. Learning to dress 3d people in generative clothing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020. 1
- [53] D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, and C. Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. volume 36, July 2017. 2, 7
- [54] D. Metaxas and D. Terzopoulos. Shape and nonrigid motion estimation through physics-based synthesis. *IEEE Trans. PAMI*, 15(6):580–591, 1993. 2
- [55] A. Mustafa, H. Kim, J.-Y. Guillemaut, and A. Hilton. General dynamic scene reconstruction from multiple view video. In *ICCV, 2015*. 1
- [56] R. Natsume, S. Saito, Z. Huang, W. Chen, C. Ma, H. Li, and S. Morishima. Siclope: Silhouette-based clothed people. *arXiv preprint arXiv:1901.00049*, 2018. 2
- [57] R. A. Newcombe, D. Fox, and S. M. Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 2
- [58] S. Orts-Escolano, C. Rhemann, S. Fanello, W. Chang, A. Kowdle, Y. Degtyarev, D. Kim, P. L. Davidson, S. Khamis, M. Dou, et al. Holoportation: Virtual 3d teleportation in real-time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 741–754. ACM, 2016. 2

- [59] S. I. Park and J. K. Hodgins. Data-driven modeling of skin and muscle deformation. In *ACM Transactions on Graphics (TOG)*, volume 27, page 96. ACM, 2008. 2
- [60] C. Patel, Z. Liao, and G. Pons-Moll. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020. 1, 2
- [61] G. Pavlakos, V. Choutas, N. Ghorbani, T. Bolkart, A. A. Osman, D. Tzionas, and M. J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2
- [62] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *CVPR*, 2018. 2
- [63] R. Plänkers and P. Fua. Tracking and modeling people in video sequences. *Computer Vision and Image Understanding*, 81(3):285–302, 2001. 2
- [64] G. Pons-Moll, S. Pujades, S. Hu, and M. Black. ClothCap: Seamless 4D clothing capture and retargeting. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 36(4), 2017. 1, 2
- [65] G. Pons-Moll, J. Romero, N. Mahmood, and M. J. Black. Dyna: a model of dynamic human shape in motion. *ACM Transactions on Graphics (TOG)*, 34(4):120, 2015. 1, 2
- [66] F. Prada, M. Kazhdan, M. Chuang, A. Collet, and H. Hoppe. Spatiotemporal atlas parameterization for evolving meshes. *ACM Transactions on Graphics (TOG)*, 36(4):58, 2017. 2
- [67] A. Pumarola, J. Sanchez-Riera, G. P. T. Choi, A. Sanfeliu, and F. Moreno-Noguer. 3dpeople: Modeling the geometry of dressed humans. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 1
- [68] H. Rhodin, N. Robertini, D. Casas, C. Richardt, H.-P. Seidel, and C. Theobalt. General automatic human shape and motion capture using volumetric contour cues. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *ECCV*, pages 509–526, Cham, 2016. Springer International Publishing. 2
- [69] J. Romero, D. Tzionas, and M. J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):245:1–245:17, Nov. 2017. 2
- [70] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. *CoRR*, abs/1905.05172, 2019. 1, 2, 6, 7, 8
- [71] L. Sigal, S. Bhatia, S. Roth, M. J. Black, and M. Isard. Tracking loose-limbed people. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 1, pages I–421. IEEE, 2004. 2
- [72] T. Simon, H. Joo, I. Matthews, and Y. Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017. 3, 5
- [73] M. Slavcheva, M. Baust, D. Cremers, and S. Ilic. Killing-fusion: Non-rigid 3d reconstruction without correspondences. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 3, page 7, 2017. 2
- [74] C. Sminchisescu and B. Triggs. Estimating articulated human motion with covariance scaled sampling. *The International Journal of Robotics Research*, 22(6):371–391, 2003. 2
- [75] O. Sorkine and M. Alexa. As-rigid-as-possible surface modeling. In *Proceedings of the Fifth Eurographics Symposium on Geometry Processing, SGP '07*. Eurographics Association, 2007. 5
- [76] J. Starck and A. Hilton. Surface capture for performance-based animation. *IEEE Computer Graphics and Applications*, 27(3):21–31, 2007. 3
- [77] R. W. Sumner, J. Schmid, and M. Pauly. Embedded deformation for shape manipulation. *ACM Trans. Graph.*, 26(3), July 2007. 3, 5
- [78] Y. Tao, Z. Zheng, Y. Zhong, J. Zhao, D. Quionhai, G. Pons-Moll, and Y. Liu. Simulcap : Single-view human performance capture with cloth simulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, jun 2019. 2
- [79] Treedys. <https://www.treedys.com/>. 3
- [80] G. Varol, D. Ceylan, B. Russell, J. Yang, E. Yumer, I. Laptev, and C. Schmid. BodyNet: Volumetric inference of 3D human body shapes. In *ECCV*, 2018. 2
- [81] D. Vlastic, I. Baran, W. Matusik, and J. Popović. Articulated mesh animation from multi-view silhouettes. In *ACM Transactions on Graphics (TOG)*, volume 27, page 97. ACM, 2008. 1, 3
- [82] D. Vlastic, P. Peers, I. Baran, P. Debevec, J. Popović, S. Rusinkiewicz, and W. Matusik. Dynamic shape capture using multi-view photometric stereo. *ACM Transactions on Graphics (TOG)*, 28(5):174, 2009. 1
- [83] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. Convolutional Pose Machines. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3, 5
- [84] X. Wei, P. Zhang, and J. Chai. Accurate realtime full-body motion capture using a single depth camera. *ACM TOG (Proc. SIGGRAPH Asia)*, 31(6):188:1–188:12, 2012. 2
- [85] A. Weiss, D. Hirshberg, and M. J. Black. Home 3d body scans from noisy image and range data. In *Proc. ICCV*, pages 1951–1958. IEEE, 2011. 2
- [86] C. Wu, C. Stoll, L. Valgaerts, and C. Theobalt. On-set Performance Capture of Multiple Actors With A Stereo Camera. In *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia 2013)*, volume 32, pages 161:1–161:11, November 2013. 1, 3
- [87] C. Wu, K. Varanasi, and C. Theobalt. Full body performance capture under uncontrolled and varying illumination: A shading-based approach. In *ECCV*, pages 757–770, 2012. 3
- [88] D. Xiang, H. Joo, and Y. Sheikh. Monocular total capture: Posing face, body, and hands in the wild. *CoRR*, abs/1812.01598, 2018. 2
- [89] W. Xu, A. Chatterjee, M. Zollhöfer, H. Rhodin, D. Mehta, H.-P. Seidel, and C. Theobalt. Monoperfcap: Human performance capture from monocular video. *ACM Trans. Graph.*, 37(2):27:1–27:15, May 2018. 1, 3

- [90] J. Yang, J.-S. Franco, F. Hétroy-Wheeler, and S. Wuhrer. Estimation of Human Body Shape in Motion with Wide Clothing. In *European Conference on Computer Vision 2016*, Amsterdam, Netherlands, Oct. 2016. [2](#)
- [91] G. Ye, Y. Liu, N. Hasler, X. Ji, Q. Dai, and C. Theobalt. Performance capture of interacting characters with handheld kinects. In *ECCV*, volume 7573 LNCS, pages 828–841, 2012. [2](#), [3](#)
- [92] M. Ye and R. Yang. Real-time simultaneous pose and shape estimation for articulated objects using a single depth camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2345–2352, 2014. [2](#)
- [93] T. Yu, K. Guo, F. Xu, Y. Dong, Z. Su, J. Zhao, J. Li, Q. Dai, and Y. Liu. Bodyfusion: Real-time capture of human motion and surface geometry using a single depth camera. In *The IEEE International Conference on Computer Vision (ICCV)*. ACM, October 2017. [2](#)
- [94] T. Yu, Z. Zheng, K. Guo, J. Zhao, Q. Dai, H. Li, G. Pons-Moll, and Y. Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *The IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2018. [2](#)
- [95] C. Zhang, S. Pujades, M. Black, and G. Pons-Moll. Detailed, accurate, human shape estimation from clothed 3D scan sequences. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. Spotlight. [2](#)
- [96] P. Zhang, K. Siu, J. Zhang, C. K. Liu, and J. Chai. Leveraging depth cameras and wearable pressure sensors for full-body kinematics and dynamics capture. *ACM Transactions on Graphics (TOG)*, 33(6):14, 2014. [2](#)
- [97] Q. Zhang, B. Fu, M. Ye, and R. Yang. Quality dynamic human body modeling using a single low-cost depth camera. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 676–683. IEEE, 2014. [2](#)
- [98] Z. Zheng, T. Yu, H. Li, K. Guo, Q. Dai, L. Fang, and Y. Liu. HybridFusion: Real-Time Performance Capture Using a Single Depth Sensor and Sparse IMUs. In *Proceedings of the 15th European Conference on Computer Vision*, Munich, Germany, Sept. 2018. Computer Vision Foundation. [2](#)
- [99] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu. Deephuman: 3d human reconstruction from a single image. *CoRR*, abs/1903.06473, 2019. [1](#), [2](#), [6](#), [7](#), [8](#)
- [100] M. Zollhöfer, M. Nießner, S. Izadi, C. Rhemann, C. Zach, M. Fisher, C. Wu, A. Fitzgibbon, C. Loop, C. Theobalt, and M. Stamminger. Real-time non-rigid reconstruction using an rgb-d camera. *ACM Transactions on Graphics (TOG)*, 33(4), 2014. [3](#)