NerfCap: Human Performance Capture with Dynamic Neural Radiance Fields

Kangkan Wang, Sida Peng, Xiaowei Zhou, Jian Yang, and Guofeng Zhang

Abstract—This paper addresses the challenge of human performance capture from sparse multi-view or monocular videos. Given a template mesh of the performer, previous methods capture the human motion by non-rigidly registering the template mesh to images with 2D silhouettes or dense photometric alignment. However, the detailed surface deformation cannot be recovered from the silhouettes, while the photometric alignment suffers from instability caused by appearance variation in the videos. To solve these problems, we propose NerfCap, a novel performance capture method based on the dynamic neural radiance field (NeRF) representation of the performer. Specifically, a canonical NeRF is initialized from the template geometry and registered to the video frames by optimizing the deformation field and the appearance model of the canonical NeRF. To capture both large body motion and detailed surface deformation, NerfCap combines linear blend skinning with embedded graph deformation. In contrast to the mesh-based methods that suffer from fixed topology and texture, NerfCap is able to flexibly capture complex geometry and appearance variation across the videos, and synthesize more photo-realistic images. In addition, NerfCap can be pre-trained end to end in a self-supervised manner by matching the synthesized videos with the input videos. Experimental results on various datasets show that NerfCap outperforms prior works in terms of both surface reconstruction accuracy and novel-view synthesis quality.

Index Terms—Human performance capture, dynamic neural radiance fields, human deformation fields.

1 Introduction

Human performance capture is an important problem in computer vision and graphics that has a variety of applications such as telepresence, film production, and sports broadcasting. From a multi-view or monocular video of a human performer in general daily clothing, the goal is to reconstruct a dynamic 3D model sequence of the performer with space-time coherent geometry and appearance that enable photo-realistic rendering of the performer. This is an extremely challenging problem as the human motion could be arbitrarily complex with non-rigid surface deformations and the appearance may be temporally varying from frame to frame due to geometry and illumination change.

Previous human performance capture systems either rely on well-controlled multi-camera studios [1], [2], [3], [4], [5] or require depth sensors [6], [7], [8], [9], [10] for high-quality 3D reconstruction. The complicated hardware of multi-view setups or depth sensors prohibit these systems from being widely deployed for daily applications. On the contrary, monocular human performance capture methods [11], [12], [13] leverage observations in color images, e.g., body joints and silhouettes, and non-rigidly fit an off-line reconstructed template model to these observations. But detailed surface deformation cannot be recovered from these sparse features. LiveCap [12] tracks the deforming surface using a dense photometric constraint between the projected template color and the pixel color. The template textures are fixed and usually cannot match the observed images due to illumination change or self-shadowing, making the tracking fail. In addition, it is difficult to render photo-realistic videos without modeling the temporally-varying appearance.

Fig. 1. Given a sparse-view or monocular video of a performer in general clothing, NerfCap captures the space-time coherent geometry and appearance that enable photo-realistic rendering of the performer. For each example in this figure, we use four views for performance capture and show the ground-truth image in a novel view, the recovered geometry (overlayed on the image), and the novel-view synthesis result.

In this paper, we propose a novel performance capture method named NerfCap by representing the 3D performer as a neural radiance field (NeRF) [14]. Compared against polygon meshes, NeRFs can more flexibly represent temporally varying geometry and appearance without being concerned about the topological change, while being able to render more photo-realistic images with neural volume rendering [15], [16], [17], [18], [19]. Specifically, we first learn a canonical NeRF from the given 3D template model of the performer. To enable the NeRF model to track detailed human motion in the videos, we equip the NeRF with a de-

- *K. Wang, and J. Yang are with the Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, P.R. China. E-mail: {wangkangkan, csjyang}@njust.edu.cn.*
- *S. Peng, X. Zhou, and G. Zhang are with the State Key Lab of CAD&CG, Zijingang Campus, Zhejiang University, Hangzhou 310058, P.R. China. E-mail: {pengsida, xwzhou}@zju.edu.cn, zhangguofeng@cad.zju.edu.cn.*
- *Corresponding author: Guofeng Zhang.*

formation field and optimize both the canonical NeRF and the deformation field by comparing the rendered images with the input video frames, so that both the temporally varying geometry and appearance can be captured.

To represent the deformation field, previous methods either adopt the linear blend skinning [15], [16] that cannot capture nonlinear local deformation or combine the linear blend skinning with a residual displacement [20], [21] that is unconstrained by any motion prior. These related works do not adopt any motion prior in the human NeRF to effectively model the motion of loose clothes (e.g., dresses). Without a motion prior to regularize the dense motion field, the joint optimization of the canonical NeRF and deformation fields forms an under-constrained problem [18], [19]. Thus, these methods mainly handle dynamic humans with tight clothes since the clothing motion is small, but may fail to accurately reconstruct the local deformations of loose clothes that have arbitrary and complex deformations. Instead, we propose to model the deformation field as the sum of skeleton-driven deformation and the non-rigid surface deformation regularized by embedded graph, so that we are able to capture both large body movement and detailed surface deformation during the performance. In our deformation model, we explicitly represent the nonlinear local deformation by utilizing the embedded graph prior defined on the template surface, so local deformations on moving clothes can be recovered reliably. In addition, the prior works [15], [16], [20], [21] are primarily proposed to synthesize novel views for dynamic humans, while our goal is to not only achieve high-quality novel-view synthesis but also capture the dense space-time coherent geometry with frame-to-frame correspondences by integrating the human template tracking into the dynamic NeRF. The experimental results on DynaCap [22], DeepCap [13], MIT datasets [4] and our own captured dataset demonstrate that the proposed method can accurately capture human performance and synthesize photo-realistic novel views of the performer from a sparse-view or monocular video.

In summary, the main contributions of this work are:

- A novel human performance capture method NerfCap that is able to accurately recover temporally-varying geometry and appearance of the performer. The dense geometric and photometric constraints presented in the dynamic human NeRF obviously improve both the accuracy and robustness of the human template tracking compared to the state-of-the-arts [12], [13].
- A dynamic NeRF model that combines the embedded graph deformation with the linear blend skinning to represent complex deformation of humans in general clothing. The embedded graph prior is incorporated in the dynamic human NeRF to effectively model nonlinear local deformations of loose clothes which is not handled in the prior works.
- Our framework is trained in an end to end manner which takes advantages of the input video frames as the supervision data.
- The state-of-the-art performance on DynaCap [22], DeepCap [13] and MIT datasets [4] in terms of surface reconstruction and view synthesis.

2 Related Work

2.1 Template-free human performance capture

Template-free methods [6], [7] based on depth sensors obtain impressive human reconstruction in realtime by volumetrically fusing captured depth maps, but they only works reliably for small and slow motions. Some approaches [8], [9], [10] increase tracking robustness for large human motions by integrating a skeletal motion prior or a parametric model which shows the significance of motion prior in performance capture. Recent works [23], [24] adopt deep neural networks to recover 3D humans from depth scans. For example, SCANimate [24] presents an end-to-end trainable framework that creates an animatable avatar without template surface registration. However, these methods reconstruct 3D humans from depth scans and cannot synthesis photo-realistic videos. Also, depth-based methods do not directly generalize to the video-based setting because of high power consumption and data capture limitations (e.g., capture distance and sensitive to sunlight). With the progress in deep learning, many template-free approaches estimate 3D detailed human shapes from color images through implicit function [25], [26], hierarchical mesh deformation [27], normal map refinement [28], or UV mapping [29]. Although these methods can recover the surface details, the recovered shapes lack of semantic information that parameterizes human pose, shape and clothing. Also, they do not recover per-vertex correspondences between different frames which limits the applications like video editing and augmented reality. Recent works [30], [31] reconstruct a personalized high-fidelity avatar from a monocular video, which is restricted to self-rotating humans with A-pose.

2.2 Template-based human performance capture

Many deep learning based methods directly learn a parametric model (e.g., the SMPL model [32]) from a single color image [33], [34], [35], [36], [37], [38], point clouds [39], [40], [41], or videos [42], [43], [44], [45] through convolutional neural networks. These methods can yield good performance on recovering the naked body shape under clothing but fail to reconstruct clothing details because parametric models do not represent garments. Some works try to parameterize the clothed humans through an extra displacement layer [46], [47], [48] or by modeling the clothing explicitly [49], [50]. However, it is difficult for these methods to deal with loose clothes or predict realistic clothing deformations.

Template mesh-based human performance capture employs a template mesh as prior and exploits multi-view geometry to track the motion of a person. Typical approaches [11], [12], [13] align a person-specific template model to the 2D observations using non-rigid registration. With a person-specific template mesh, these methods can achieve realistic performance capture, and recover semantic information of the humans, and obtain space-time per-vertex correspondences. DeepCap [13] estimates the non-rigid surface deformation of the template mesh with weak supervision of multi-view joints and silhouettes. But only with these sparse features, they cannot accurately recover the detailed surface deformation of the human. LiveCap [12] combines dense photometric constraints with silhouette

constraints, but their template with fixed texture cannot express appearance variations in the videos. In this work, the dynamic NeRF flexibly represents the temporally-varying geometry and appearance of the performer, enabling our method recover surface geometry accurately and render photo-realistic videos from novel views.

2.3 Neural implicit representation-based methods

The 3D surfaces can be reconstructed by learning implicit representations such as voxel representation [28], [51], implicit function [25], [26], or neural radiance fields [14], [16], [17], [18], [19]. Voxel representation is memory intensive which restricts its ability to produce fine-scale detailed surfaces. The implicit function based methods can produce high-resolution surfaces, but the human limbs are often missing due to occlusion or difficult poses. Neural radiance fields (NeRF) [14] represents scenes with implicit fields of density and color and can achieve photo-realistic synthesis. NeRF is extended for dynamic scenes by jointly optimizing a canonical NeRF and a set of deformation fields represented as translational vector field [17], [18] or SE(3) field [19]. To regularize the dense motion field, these methods adopt an as-rigid-as-possible constraint [18], [19], and obtain good performance on small deformations. However, they are not adaptable for dynamic humans because without human motion prior jointly optimizing NeRF with the deformation field is an extremely under-constrained problem.

Neuralbody [15] and AniNeRF [16] represent a dynamic human NeRF based on the SMPL model [32] or by combining skeleton-driven deformation [52] with learned blend weights, which regularizes the learning of deformation fields and achieves impressive novel-view synthesis. However, their performance degrades dramatically on the humans in loose clothes because the linear bend skinning cannot capture nonlinear local deformation of moving clothes. The extended work, AniSDF [53] of AniNeRF [16] utilizes signed distance fields to model the human geometry, and achieves better geometry reconstruction accuracy compared to the density fields [16]. However, with the same limitation of AniNeRF, AniSDF [53] cannot accurately reconstruct humans in loose clothes. H-NeRF [54] unifies NeRF and signed distance field for recovering dynamic humans, which adopts the articulated deformation model similar to [15], [16]. The other kind methods of human NeRF [20], [21] represent nonrigid deformation as a displacement on top of the linear blend skinning. However, without any constraint of motion prior, the residual displacement cannot capture the motion of loose clothes effectively, and the clothing geometry fails to be recovered under complex and arbitrary deformations. In our work, we model nonrigid deformations using an embedded deformation graph on the human template since the embedded deformation is effective to model the motion of loose clothes. In addition, motion tracking of human geometry [9], [12], [13] is to reconstruct not only the geometry of each frame but also space-time dense correspondences across the sequences, which are crucial for the downstream applications in AR/VR. However, the existing human NeRF methods [15], [16], [20], [21], [53], [54] build temporal correspondences among the video frames using the skeleton-driven framework, which fails to represent the motion of loose clothes. In contrast, our method mainly focuses on the tracking of a human template as [12], [13] and thus obtains the frame-to-frame dense correspondences.

3 Proposed Approach

Given a sparse-view or monocular video of a human in general clothing and a 3D template model of the human, our goal is to capture the densely deforming surface and synthesize free-viewpoint videos of the human. We achieve this goal with a novel method of human performance capture with dynamic neural radiance fields (NeRF) which mainly consists of two components, namely the human deformation field network and the canonical NeRF network. Figure 2 illustrates an overview of the proposed framework. The human deformation field network regresses the non-rigid deformation of the dense surface. Combined with the skeleton-driven deformation, it captures the dense deformation field between the canonical frame and video frames (Sec. 3.2). The canonical NeRF network estimates a radiance field in the canonical frame which is dynamically warped to video frames with the deformation field (Sec. 3.3). The dynamic NeRF allows us to render images of the performer from any viewpoint. By comparing the rendered images with the input video frames, the dynamic NeRF can capture the temporally-varying geometry and appearance of the moving human.

3.1 Pre-processing

We assume a rigged 3D template mesh of the performer has been reconstructed off-line, e.g., using image-based reconstruction method as in [12]. Then, a canonical NeRF of the performer is constructed from the template model by supervising the occupancy network with the template geometry. The canonical occupancy network is pre-trained and frozen in the training of other networks. During training, we randomly sample one image $I_{i,c}$, where $i \in [0, F]$ and $c \in [0, C]$ denote the frame index and camera index, respectively. For each frame $I_{i,c}$, we extract the foreground human mask using [55] and obtain the 3D human pose with [56]. In the framework overview illustrated in Fig. 2, we assume the input data have been pre-processed.

3.2 Human Deformation Field

The 3D skeleton pose alone is hardly able to model non-rigid deformations of the human in general clothing such as the swinging of a skirt. To address this issue, we disentangle the human deformation field into the non-rigid surface deformation and the articulated skeletal motion. After non-rigid surface deformation is performed on the canonical template, we apply skeleton-driven deformation to model articulated motions. This significantly simplifies the learning of non-rigid deformations by alleviating ambiguities in the movements of mesh vertices caused by pose variations.

3.2.1 Non-rigid Surface Deformation

We model non-rigid surface deformation with embedded deformation [57] that computes a warping field based on a deformation graph. An embedded deformation graph G with K nodes (K is around 500 in our experiments) is automatically built on the template. The node transformations are parameterized with Euler angles $\boldsymbol{A} \in \mathbb{R}^{K\times 3}$ and

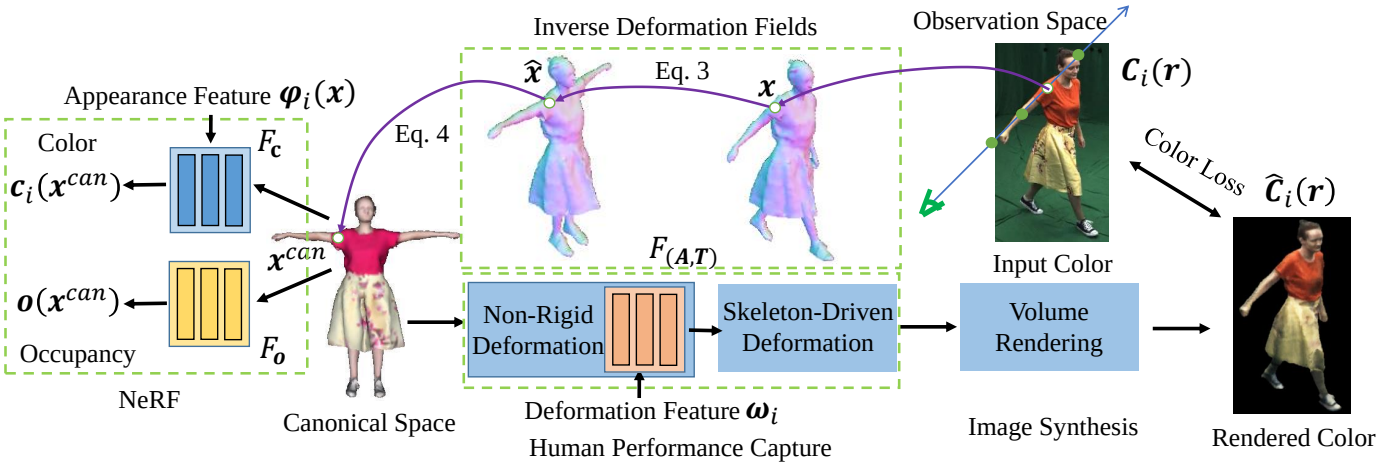


Fig. 2. Overview of the proposed framework. Our framework can capture the human performance and synthesize novel-view videos from a sparse-view or monocular video. We jointly optimize the human deformations and a canonical neural radiance field (NeRF) defined on the template model. Please refer to Sec. 3 for detailed description. To simplify the framework overview, we illustrate the extraction process of deformation features and appearance features in Fig. 3.

translations $\boldsymbol{T} \in \mathbb{R}^{K\times 3}$. For point $\boldsymbol{v}$, its new position $\boldsymbol{y}$ after the non-rigid deformation is computed as,

$$\boldsymbol{y} = \sum_{k\in\mathcal{N}(\boldsymbol{v})} w(\boldsymbol{v},\boldsymbol{g}_k)[\boldsymbol{R}(\boldsymbol{A}_k)(\boldsymbol{v}-\boldsymbol{g}_k)+\boldsymbol{g}_k+\boldsymbol{T}_k], \quad (1)$$

where $\mathcal{N}(\boldsymbol{v})$ denotes the set of neighboring nodes that influence the movement of point $\boldsymbol{v}$, $\boldsymbol{R}(\cdot):\mathbb{R}^3\rightarrow SO(3)$ converts the Euler angles to rotation matrices, and $w(\boldsymbol{v},\boldsymbol{g}_k)$ is the deformation weight of the k-th neighboring node on $\boldsymbol{v}$ which can be pre-computed [57]. The feature extraction process is illustrated in Figure 3. We extract the deformation features using ResNet50 as [13] and modify the last fully connected layer to output a $L1$-dim feature vector $\boldsymbol{\omega}_{i,c} = F_d(I_{i,c})$. The multi-view features are then aggregated with a max pooling operator to output the deformation feature of each frame $\boldsymbol{\omega}_i = \text{MaxPool}(\boldsymbol{\omega}_{i,1},...,\boldsymbol{\omega}_{i,C})$. From the deformation feature $\boldsymbol{\omega}_i$, the non-rigid deformation is estimated using an MLP network $F_{(\boldsymbol{A},\boldsymbol{T})}:\boldsymbol{\omega}_i\rightarrow(\boldsymbol{A}_i,\boldsymbol{T}_i)$. It has been demonstrated in [9], [11], [13] that the embedded deformation is effective to capture the performance of humans in daily clothes. In our method, we contribute to introduce the embedded graph prior in the dynamic NeRF to enhance its capability of representing humans in various clothes.

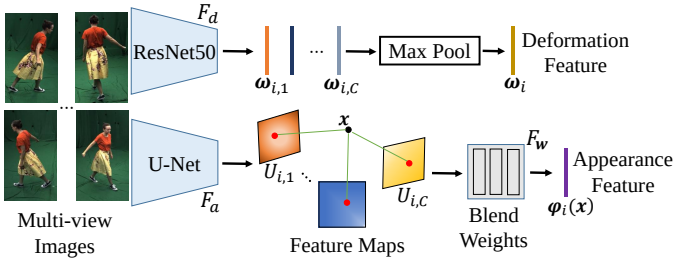


Fig. 3. Illustration of the feature extraction. The deformation features are used to regress the non-rigid deformation, while the appearance features of sampled points are input to the color model of the canonical NeRF. The extraction methods for deformation features and appearance features are described in Sec. 3.2.1 and Sec. 3.3.1, respectively.

3.2.2 Skeleton-Driven Deformation

We use the 3D human skeleton to model the articulation deformation. For each frame in the video, we obtain the SMPL model with [56]. For each vertex on the 3D template, we compute its skinning weight by finding the closest surface point on the SMPL model as [16]. Based on the non-rigidly deformed mesh using the embedded graph, we further apply the linear blend skinning [52] to obtain the deformed and posed point in the observation space as,

$$\hat{\boldsymbol{v}} = [\sum_{j=1}^{J} w(\boldsymbol{y})_j \boldsymbol{G}_j]\boldsymbol{y}, \quad (2)$$

where J is the number of joints in the human skeleton, $w(\boldsymbol{y})_j$ is the skinning weight of j-th part on $\boldsymbol{y}$, and $\boldsymbol{G}_j \in SE(3)$ denotes the rigid transformation matrices.

3.3 Human Performance Capture

In this section, we first introduce the dynamic NeRF for modeling dynamic humans and then describe the method for human performance capture with the dynamic NeRF.

3.3.1 Dynamic Neural Radiance Fields for Humans

Neural Radiance Fields (NeRF) [14] represents a static scene using a neural implicit 3D volume. The 3D volume is approximated with an MLP network $\boldsymbol{F}_\theta:(\boldsymbol{x},\boldsymbol{d})\rightarrow(\boldsymbol{c},\boldsymbol{\sigma})$ that maps a 3D position $\boldsymbol{x}\in\mathbb{R}^3$ and viewing direction $\boldsymbol{d}\in\mathbb{R}^3$ to an RGB color $\boldsymbol{c}=\boldsymbol{c}_\theta(\boldsymbol{x},\boldsymbol{d})$ and a density $\boldsymbol{\sigma}=\boldsymbol{\sigma}_\theta(\boldsymbol{x})$.

Some recent works extend the original NeRF to dynamic humans, but they are limited to small and slow motions [18], [19] or cannot recover detailed deformations for humans in loose clothes [16], [17], [20]. Here, we present a dynamic human NeRF which can recover the geometry and appearance of humans in general clothing and with large body motions.

To obtain the dynamic NeRF at frame i, we transform the sampled observation-space point $\boldsymbol{x}$ to canonical space with $\boldsymbol{x}^{can} = \boldsymbol{T}_i(\boldsymbol{x})$. The transformations $\boldsymbol{T}_i$ are obtained with the inverse deformation fields of 3D template surface. In this way, our method associates the deformation field of the canonical NeRF with the human deformation, allowing us

to optimize the canonical NeRF and capture human performance simultaneously. Specifically, for an observation-space point $\boldsymbol{x}$, we search the closest vertex $\boldsymbol{v}$ on the deformed 3D template, and transform it with the inverse linear blend skinning of $\boldsymbol{v}$:

$$\hat{\boldsymbol{x}} = [\sum_{j=1}^{J} w(\boldsymbol{v})_j \boldsymbol{G}_j]^{-1} \boldsymbol{x}. \tag{3}$$

Then, we apply the inverse graph deformation of $\boldsymbol{v}$ to $\hat{\boldsymbol{x}}$:

$$\begin{aligned} \boldsymbol{x}^{can} \quad = \quad & [\sum_{k \in \mathcal{N}(\boldsymbol{v})} w(\boldsymbol{v}, \boldsymbol{g}_k) \boldsymbol{A}_k]^{-1} \cdot \\ & [\sum_{k \in \mathcal{N}(\boldsymbol{v})} w(\boldsymbol{v}, \boldsymbol{g}_k)(\hat{\boldsymbol{x}} - \boldsymbol{g}_k - \boldsymbol{T}_k + \boldsymbol{A}_k \boldsymbol{g}_k)], \end{aligned} \tag{4}$$

where the notations are the same with Eq. (1). The correspondence finding based on the closest point is iterative similar to the traditional deformation process [57]. Through the two-step inverse deformation, we obtain the canonical-space point $\boldsymbol{x}^{can}$ for the observation-space point $\boldsymbol{x}$.

The occupancy and color are predicted for $\boldsymbol{x}^{can}$ based on the canonical NeRF. Instead of using the volume density in the original NeRF [14], we represent the geometry with an occupancy network [58], which can represent high-quality geometry. We pre-train the canonical occupancy network with the supervision of the template geometry and recover the temporally-varying geometry by transforming it to observation frames with the human deformation field. The canonical occupancy model $F_{\boldsymbol{o}}$ is defined as,

$$\boldsymbol{o}(\boldsymbol{x}^{can}) = F_{\boldsymbol{o}}(\gamma_{\boldsymbol{x}}(\boldsymbol{x}^{can})), \tag{5}$$

where $\gamma_{\boldsymbol{x}}$ is the positional encoding [14] for the spatial location. The positional encoding allows the network to better capture high frequency signals. To encode the temporally-varying factors such as illumination variation and exposure, we extract the appearance feature $\boldsymbol{\varphi}_i(\boldsymbol{x})$ for each sampled point $\boldsymbol{x}$ in observed frame i. The process of this feature extraction is shown in Figure 3. Given the input image $I_{i,c} \in \mathbb{R}^{H \times W \times 3}$, we use a U-Net network F_a to extract dense features $U_{i,c} \in \mathbb{R}^{H \times W \times L2}$ representing local image appearance. For the spatial point $\boldsymbol{x}$, we project it into view c at $\boldsymbol{q}$ and fetch the corresponding feature vector $U_{i,c}(\boldsymbol{q})$. The aggregated feature of $\boldsymbol{x}$ is then computed as a weighted summation of image features as $\boldsymbol{\varphi}_i(\boldsymbol{x}) = \sum_{c=1}^{C} w_c(\boldsymbol{x}) U_{i,c}(\boldsymbol{q})$ over all views. Similar to IBRNet [59], we use a MLP network to estimate the blending weights $w_c(\boldsymbol{x})$ as:

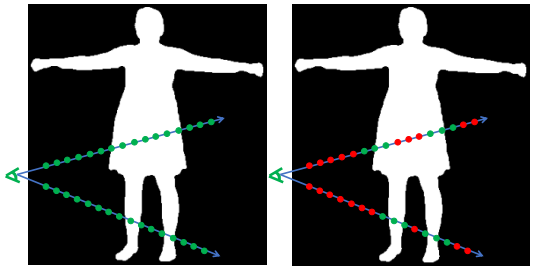
$$w_c(\boldsymbol{x}) = F_{\boldsymbol{w}}(d_c(\boldsymbol{x}), U_{i,c}(\boldsymbol{q})), \tag{6}$$

where $F_{\boldsymbol{w}}$ denotes the MLP network, and $d_c(\boldsymbol{x})$ is ray direction of $\boldsymbol{x}$ to view c. The canonical color model $F_{\boldsymbol{c}}$ is defined as,

$$\boldsymbol{c}_i(\boldsymbol{x}^{can}) = F_{\boldsymbol{c}}(\gamma_{\boldsymbol{x}}(\boldsymbol{x}^{can}), \gamma_{\boldsymbol{d}}(\boldsymbol{d}), \boldsymbol{\varphi}_i(\boldsymbol{x})), \tag{7}$$

where $\gamma_{\boldsymbol{d}}$ is the positional encoding for viewing direction.

Finally, the rendered color is obtained by integrating all sampled points in the ray with the volume rendering technique [58]. In [16], all the sampled points along the ray within a human bounding box are contributed to the volume rendering (as illustrated in Figure 4 (a)). We find it leads to obvious artifacts on the synthesized images in the experiments (shown in Figure 12). This is because observation-space points far from the human surface (red points in the Figure 4) may be deformed to points near the canonical template and fetch spurious occupancy. Actually, only the space points close to the deformed mesh (green points in the Figure 4) should have positive effects on the rendering process. Based on this fact, we take advantage of the deformed template mesh to guide the point sampling (as illustrated in Figure 4 (b)). Thus, the rendered color for ray $\boldsymbol{r}$ at frame i is define as:



(a) (b)

Fig. 4. Illustration of the mesh-guided point sampling. (a) The volume rendering of [16] integrates all the space points along the ray. (b) Our method only utilizes the points near the surface of the deformed template mesh (green points). Our method can obtain higher quality of novel-view synthesis as demonstrated in the ablation study.

$$\hat{\boldsymbol{C}}_i(\boldsymbol{r}) = \sum_{l=1}^{N} \boldsymbol{m}(\boldsymbol{x}_l) \boldsymbol{o}(\boldsymbol{x}_l^{can}) \prod_{j<l} (1 - \boldsymbol{m}(\boldsymbol{x}_j) \boldsymbol{o}(\boldsymbol{x}_j^{can})) \boldsymbol{c}_i(\boldsymbol{x}_l^{can}), \tag{8}$$

where N is the total number of sampled points on the ray, and $\boldsymbol{m}(\boldsymbol{x}_l)$ is a mask ($\boldsymbol{m}(\boldsymbol{x}_l) = 1$ if the distance of $\boldsymbol{x}_l$ is less than a threshold ε to the deformed mesh, otherwise $\boldsymbol{m}(\boldsymbol{x}_l) = 0$).

3.3.2 *Performance Capture with Dynamic NeRF*

The dynamic NeRF is obtained in the observation frame by deforming the canonical NeRF with the per-frame deformation fields of the human. From the dynamic NeRF, we are able to synthesize pixel color from free viewpoints through the volume rendering. By minimizing the error between the rendered pixel color $\hat{\boldsymbol{C}}_i(\boldsymbol{r})$ and the observed pixel color $\boldsymbol{C}_i(\boldsymbol{r})$, we enable the dynamic NeRF to capture the temporally-varying geometry and appearance of the human. Formally, the parameters of the color model $F_{\boldsymbol{c}}$, the deformation network $F_{(\boldsymbol{A},\boldsymbol{T})}$, feature extracting networks F_d and F_a, and blend weight network $F_{\boldsymbol{w}}$ are jointly optimized over the video:

$$L_{rgb} = \sum_{\boldsymbol{r} \in \boldsymbol{R}} \| \hat{\boldsymbol{C}}_i(\boldsymbol{r}) - \boldsymbol{C}_i(\boldsymbol{r}) \|_2, \tag{9}$$

where $\boldsymbol{R}$ is the set of casted rays.

To enforces local smoothness of the deforming surface, we impose an as-rigid-as-possible loss [57] on the non-rigid deformation:

$$L_{arap} = \sum_{\boldsymbol{g}_i} \sum_{\boldsymbol{g}_j \in \mathcal{N}(\boldsymbol{g}_i)} w(\boldsymbol{g}_i, \boldsymbol{g}_j) \| d_{i,j}(\boldsymbol{A}, \boldsymbol{T}) \|_2, \tag{10}$$

where

$$d_{i,j}(\boldsymbol{A}, \boldsymbol{T}) = \boldsymbol{A}_j(\boldsymbol{g}_i - \boldsymbol{g}_j) + \boldsymbol{g}_j + \boldsymbol{T}_j - (\boldsymbol{g}_i + \boldsymbol{T}_i).$$

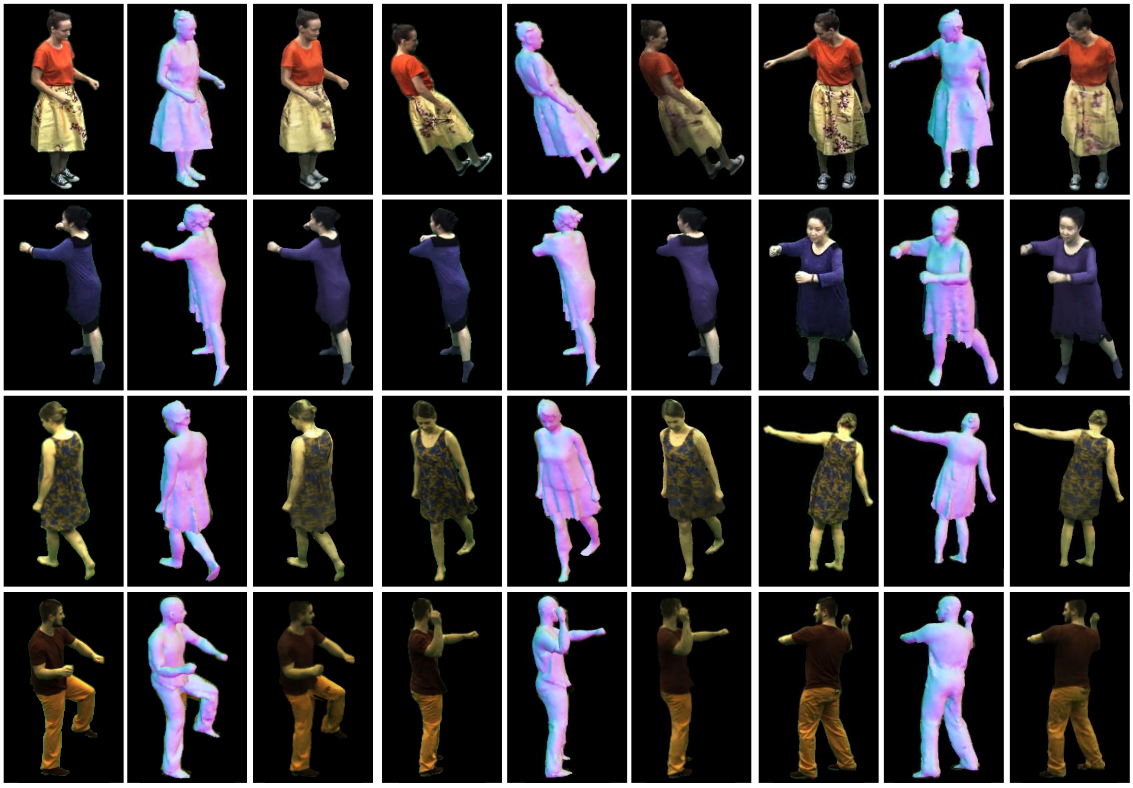


Fig. 5. Qualitative results of our method on different humans with varying types of apparel. For each example, we show the ground truth image in a novel view, the recovered geometry (overlayed on the image), and the novel-view synthesis result. From top to bottom: "FranziRed" and "LingPurple" from DynaCap dataset [22], and the subject S4 and S1 from DeepCap dataset [13]. Note that the recovered geometry precisely overlays to the image and the novel-view synthesis is photo-realistic.

Then, the total loss of our method is defined as:

$$L = L_{rgb} + \lambda L_{arap}, \tag{11}$$

where λ is a balance weight. The intuition of our method is that the learning of the NeRF and human deformations can benefit each other during the joint optimization. By minimizing the error between the rendered color and observed color (defined in Eq. 9), observed points on the human body are matched to the corresponding vertices on the template in terms of both geometry and color. The canonical NeRF is dynamically warped to the observation space with the estimated human deformation. Thus, dense geometric and photometric constraints are formed on the human deformation estimation during the learning of NeRF. With the strong ability of representing photo-realistic colors, the temporally-varying appearance and geometry of the human can be effectively captured in the dynamic NeRF.

3.4 Fast Fine-tuning

Our framework can be trained in a self-supervised manner by matching the synthesis images with the input video frames. When tested on the sequences with unseen human poses or new lighting environment, the pre-trained model may generate unsatisfactory results due to the limited training data. To obtain more accurate results, we directly apply a fast fine-tuning on the test sequences using the pre-trained model as an initialization. As verified in the ablation study, the fast fine-tuning can obviously improve the accuracy of both human performance capture and novel-view synthesis.

4 Experiments

4.1 Datasets and metrics

We conduct experiments on the publicly challenging datasets, DynaCap [22], DeepCap [13], and MIT datasets [4]. Both DynaCap and DeepCap datasets are recorded in a calibrated multi-camera studio with green screen, providing multi-view videos of 4 or 5 actors doing various actions. The videos in the DynaCap dataset are recorded with 50 to 101 synchronized and calibrated cameras at a resolution of 1285×940. The DeepCap dataset contains between 11 and 14 camera views at a resolution of 1024×1024. 3D template of all actors with texture are also provided in both datasets. We uniformly sample 15 camera views from DynaCap dataset and use all views of DeepCap dataset. We pre-train the networks on their training dataset about $20,000$ frames in total, and fine-tune on the test sequences. Note that the test data are not included in the training dataset. Four cameras around the performer are used for training and the remaining camera is selected for test. MIT dataset [4] provides the per-frame mesh in a sequence, and we use them to compare the 3D model reconstruction accuracy. We also capture multi-view videos in our motion capture system, and apply our method on the videos to recover the human performance. Qualitative results of our method on different persons are shown in Figure 5. More experimental results can be found in the supplementary video.

To evaluate the accuracy of silhouette alignment, we compute the intersection over union (IoU) (%) between the ground truth foreground mask and the 2D projection of the

estimated shape. Following [13], we evaluate the IoU on all views, on all views except the input view, and on the input view which are referred to as AMVIoU, RVIoU and SVIoU, respectively. Following [16], we evaluate the novel-view image synthesis using two metrics: peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM). For evaluation of 3D surface reconstruction, we compute mean average vertex error (MAVE) [41] in millimeter between vertices on the recovered models and the corresponding vertices on the ground truth models.

4.2 Implementation details

Our canonical NeRF implementation closely follows the original one [14], and we only use the single level NeRF. In the occupancy network, we add an occupancy layer defined in [58] at the end of the density network of [14] which maps the occupancy value into $[0, 1]$. The occupancy network is pre-trained by sampling points around the surface similar to [25], and frozen in the training of other networks. We use the same color network as [16] without outputting the density and extract the appearance features from input images instead of the per-frame latent code [16]. We use a non-rigid deformation network with depth 8, hidden size 1024, and a skip connection at the 4th layer. The blend weight network composes of three fully-connected layers with $(131, 64, 32, 1)$ neurons. The size of deformation features L_1 and size of appearance features L_2 both have dimensions of 128. The loss weight λ is set to 0.005 empirically. The Adam optimizer [60] is adopted for the training. The learning rate starts from $5e^{-4}$ and decays exponentially to $5e^{-5}$ along the optimization. The experiments are conducted on a 2080 Ti GPU. The training time of our network is about 3 days on four-view videos about $20,000$ frames. The time of the fine-tuning depends on the number of video frames and input views. For a four-view video of 300 frames with 1024×1024 resolution, the fine-tuning takes less than 1 hour. For a monocular video of 300 frames, the fine-tuning takes less than 15 minutes, which is very fast compared to the prior works [15], [16].

4.3 Comparison to state-of-the-art methods

To validate our method, we compare with three state-of-the-art methods for human body reconstruction using NeRF, Neuralbody [15], AniNeRF [16], and AniSDF [53] (the extended work of AniNeRF). We also compare with D-NeRF [17] which is proposed for dynamic scenes and applied on the human body in their paper. These methods can synthesize novel-view images, and we use their released codes for implementation. In addition, we compare with two state-of-the-art methods for human performance capture, DeepCap [13] and LiveCap [12]. Since their codes are not released, we implement them by following the implementation details and parameter settings described in their papers. The results of DeepCap [13] and LiveCap [12] are the best that we can obtain in the experiments. To make fair comparisons, we also compare our results with that of [12], [13] provided in their original video. The two methods synthesize free-viewpoint videos with a fixed texture which cannot match the varying human appearance and is not realistic as the input images. Based on the human deformation

Method	Silhouette Alignment			Image Synthesis	
	AMVIoU	RVIoU	SVIoU	PSNR	SSIM
AniNeRF [16]	88.50	87.08	90.98	21.77	84.97
D-NeRF [17]	91.66	90.36	93.93	21.84	85.93
Neuralbody [15]	89.89	89.03	91.40	22.92	88.29
AniSDF [53]	**93.37**	**92.05**	**95.67**	23.08	88.26
DeepCap [13]	86.17	85.75	86.90	21.37	85.52
LiveCap [12]	87.72	87.17	88.69	21.74	86.15
Our method	88.96	88.32	90.09	**24.52**	**90.79**

Tested on the subject S4 of DeepCap dataset [13].

Method	Silhouette Alignment			Image Synthesis	
	AMVIoU	RVIoU	SVIoU	PSNR	SSIM
AniNeRF [16]	80.89	79.45	84.87	17.10	71.82
D-NeRF [17]	91.76	91.33	**92.93**	20.08	83.44
Neuralbody [15]	89.92	89.53	90.98	21.24	87.61
AniSDF [53]	**91.78**	**91.49**	92.59	19.89	86.52
DeepCap [13]	85.38	85.37	85.41	19.38	85.39
LiveCap [12]	85.95	85.92	86.04	19.79	85.94
Our method	88.08	88.07	88.12	**23.18**	**91.49**

Tested on the "FranziRed" of DynaCap dataset [22].

TABLE 1
Comparison to state-of-the-art methods in terms of silhouette alignment and image synthesis. Note that, the silhouette IoUs of the first four methods are high because they use multi-view human masks in recovering the human geometry, and their reconstructed surfaces are bumpy and lack details as shown in Figure 6. In contrast, our method can capture the space-time coherent geometry and obtain the frame-to-frame dense correspondences across the videos.

recovered using the two methods, we synthesize novel-view images with the color model estimated by our method.

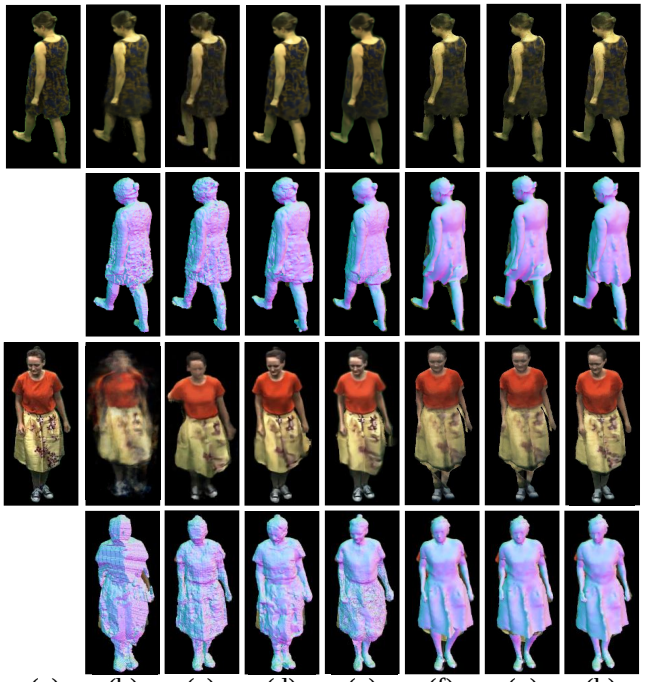


Fig. 6. Comparison to the state-of-the-arts. (a) Ground truth. (b) AniNeRF [16]. (c) D-NeRF [17]. (d) Neuralbody [15]. (e) AniSDF [53]. (f) DeepCap [13]. (g) LiveCap [12]. (h) Our method. In each example, we show the novel-view synthesis and reconstructed geometry.

4.3.1 Comparison to novel-view synthesis methods of dynamic humans [15], [16], [17], [53]

We perform both qualitative and quantitative comparisons on "FranziRed" from DynaCap and subject S4 of DeepCap. The qualitative results are shown in Figure 6 and the quantitative results are listed in Table 1. Without any motion prior, D-NeRF [17] cannot estimate the human motion accurately (e.g., the missing limb on "FranziRed') in the under-constrained problem. The performance of Neuralbody [15],

AniNeRF [16], and AniSDF [53] degrades dramatically on humans in loose clothes, and AniNeRF even cannot converge on the "FranziRed" data. AniSDF [53] performs better than AniNeRF, but they fail to recover the geometry details of loose skirts accurately. This is because the skeleton-driven deformation adopted in [15], [16], [53] cannot represent the motion of loose skirts. By combining skeleton-driven deformation with embedded graph deformation, our method can not only model large body motions but also recover the detailed surface deformations of humans in general clothing. With the robust and accurate motion prior, our method reliably estimates both the human deformation and canonical NeRF, resulting in higher quality novel-view synthesis. Instead of introducing a per-frame latent code as [15], [16], [53], our method extracts deep features from the input images, allowing us to train the networks end to end.

Using a personalized template is not a restriction in our method since a template can be obtained conveniently with many robust methods [6], [8], [9]. We do not use a template model in the implementation of the four compared methods (i.e., [15], [16], [17], and [53]) because incorporating a template needs to estimate non-rigid deformations and requires large changes in the method. Note that the methods of [15], [16], [53] use the SMPL template models. These methods obtain high accuracy of silhouette alignment because they recover the human geometry by using multi-view human masks which is similar to the process of visual hull [4]. Also, their reconstructed surfaces are bumpy and lack details as shown in Figure 6. On the contrary, our method tracks the deforming geometry of the template surface and provides the frame-to-frame correspondences.

Fig. 7. Comparisons with the original results of DeepCap [13] and LiveCap [12]. (a,d) LiveCap [12]. (b,e) DeepCap [13]. (c,f) Our method. We show the reconstructed geometry overlayed on the input view (first row) and reference view (second row). The results of DeepCap [13] and LiveCap [12] are from the original video of [13]. The comparisons show that our method captures the deforming geometry more accurately.

4.3.2 Comparison to human performance capture methods [12], [13]

DeepCap [13] is trained with the weak supervision of multi-view joints and silhouettes. With these sparse features, the detailed surface deformations cannot be recovered. In addition to silhouettes, LiveCap [12] also utilizes a photometric alignment constraint by matching the template with fixed color to pixel color on the images. Since the appearance varies in different frames and views, the photometric constraint leads to erroneous alignment. In our method, the dynamic NeRF captures the temporally-varying appearance from the input images and forms dense constraints for human deformation estimation in both geometry and appearance. Thus, our method can recover the deforming surface more accurately than these state-of-the-art methods as shown in Figure 6. In addition, we compare the frame-by-frame rendering result on the same sequence with DeepCap [13] and LiveCap [12]. Similar to [13] and [12], our results are generated on the input monocular video. The comparison results are shown in Figure 7. For the complete sequence, please refer to our supplementary video. Both [13] and [12] temporally smooth the output mesh vertices with a filter to mitigate the jitter problem, while our method recovers the smoothly deforming geometry without using any temporal filter. The comparison results demonstrate our method is more robust and accurate than [13] and [12] on human performance capture. Another significant improvement of our method is photo-realistic free-viewpoint synthesis compared to [13] and [12].

Method	AMVIoU	RVIoU	SVIoU	MAVE ↓
DeepCap [13]	84.19	84.12	84.32	17.24
LiveCap [12]	86.64	86.13	87.52	13.59
Our method	**91.15**	**90.91**	**91.58**	**11.16**

TABLE 2
Reconstruction accuracies using different methods on the "samba" data from MIT dataset [4].

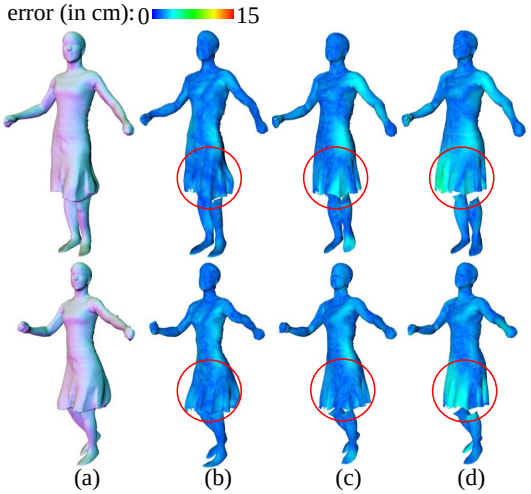


Fig. 8. The visualization of reconstruction accuracies using different methods on the "samba" data from MIT dataset [4]. (a) Input image. (b) Our method. (c) LiveCap [12]. (d) DeepCap [13]. Our method can recover detailed surface deformations more accurately (e.g., the skirt in this example).

4.4 Comparison to [13] and [12] in terms of surface reconstruction accuracy

To further demonstrate the superiority of our method over the state-of-the-art methods for human performance capture, i.e., DeepCap [13] and LiveCap [12], we perform the comparison in terms of 3D model reconstruction accuracy. We choose the "samba" data in loose skirt (about 140 frames) from MIT dataset [4]. This dataset is very challenging with complex clothing motions of swinging skirt. We synthesize videos of 14 views using the same camera parameters of DeepCap dataset [13]. Since there is no texture for

Method	AMVIoU	PSNR	SSIM
1 camera view	88.07	21.10	86.44
2 camera views	88.43	23.34	89.81
6 camera views	89.02	**24.81**	**91.10**
50 video frames	92.06	23.99	90.26
150 video frames	**95.63**	24.53	90.70
600 video frames	88.74	24.32	90.51
1500 video frames	87.99	23.85	89.25
w/o nonrigid deformation	81.54	21.10	86.72
w/o skeleton deformation	2.32	16.07	75.82
w/o fast fine-tuning	87.14	22.60	88.37
w/o mesh-guided sampling	88.47	23.96	90.34
with human NeRF in [21]	84.47	24.04	90.60
with human NeRF in [20]	85.33	24.10	90.81
ours+silhouette	88.60	24.36	90.59
ours(4 views, 300 frames)	88.96	24.52	90.79

TABLE 3
Ablation studies on S4 of DeepCap dataset [13].

the meshes, we render the normal-colored meshes as [23]. The synthetic data has the ground-truth 3D model, so the reconstruction error (MAVE in mm) of recovered 3D models can be computed as [41]. We also report the intersection over union (IoU in %) between the ground truth and rendered masks. The quantitative results are listed in Table 2, and two comparison examples with heat map accuracy are shown in Figure 8. The skirt silhouettes change dramatically under complex motions of swinging skirt, which makes the template tracking fail and leads to low recovery accuracy of skirt parts using the two compared methods. LiveCap [12] performs better than DeepCap [13] with the dense photometric constraint. However, their template with fixed color cannot match the varying appearance on different images, which causes instable tracking in the video. In contrast, our method can achieve more detailed deformations and higher reconstruction accuracy than DeepCap [13] and LiveCap [12] by utilizing the dense geometry and appearance constraints represented in the dynamic NeRF.

4.5 Ablation studies

We conduct ablation studies on S4 of DeepCap dataset [13] and report the quantitative results in Table 3.

The number of input camera views. We investigate the performances of our method under different number of camera views. We report the errors with 1, 2, 4, and 6 input views. Figure 9 shows two examples of novel-view synthesis using different numbers of camera views. We can see our method performs better in both the performance capture and novel-view synthesis with the increasing number of camera views. The human performance can be captured more completely with more input views, resulting in higher quality of novel-view synthesis.

Impact of the video length. We test the performance of our networks on videos with different numbers of frames. Figure 10 shows the comparison example. Both quantitative and qualitative comparisons show that training on the video helps the joint learning of the canonical NeRF and human deformations. But, we can see that the accuracy of the network may decrease when fitting on very long videos. Different from modeling static scenes with NeRF [59], our NeRF model is trained by using multi-view videos as the input which covers large viewpoint changes, and the moving humans have complex deformations, posing tremendous challenges on fitting very long videos.

Fig. 9. Comparison with different numbers of camera views. The artifacts are marked in cycle on the novel-view synthesis.

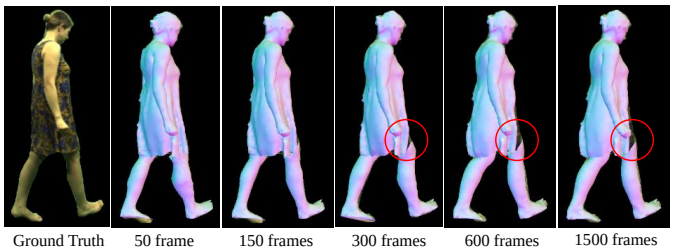


Fig. 10. Comparison with different numbers of video frames. The recovery accuracy may decrease when fitting on very long videos.

Human deformation model. We compare our method with another two cases, i.e., without non-rigid surface deformation and without skeleton-driven deformation, by removing the corresponding module in our framework. The comparison examples are shown in Figure 11. The results show that removing either of the two modules will lead to a dramatic decrease in the accuracy. Skeletal deformation alone fails to capture local nonlinear deformations, while the optimization only using non-rigid deformation is prone to local minima without skeletal motion prior. This experiment demonstrates that the human motion prior which combines embedded graph deformation and skeleton-driven deformation is essential to model the NeRF successfully for dynamic humans in loose clothes.

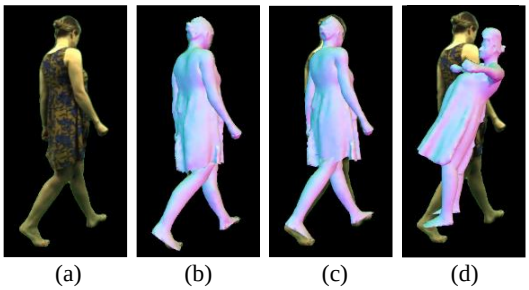


Fig. 11. Comparison with different human deformation models. (a) The ground truth image. (b) Our method. (c) Without non-rigid surface deformation. (d) Without skeleton-driven deformation.

Mesh-guided point sampling. The volume rendering method adopted in [16] use all the sampled points along the ray (similar to the original method in NeRF [14]). With the guidance of the deformed template mesh, our method efficiently considers the points near the deformed

Fig. 12. Comparison of two volume rendering methods. (a,d) The ground truth image in a novel view. (b,e) The novel-view synthesis result using our mesh-guided point sampling. (c,f) The novel-view synthesis result using the original point sampling [16]. Our method can effectively eliminate the serious artifacts (marked in cycle) by excluding the spurious points on the ray during the rendering.

mesh (within a distance threshold). We compare the two rendering methods both quantitatively and qualitatively. Two comparison examples are shown in Figure 12. We can see that there are clear artifacts in the synthesized images using the rendering method of [16]. Some sampled points may obtain spurious occupancy when transformed to the surface of the canonical template, and affect the rendering of the pixel color during point integration along the ray. In contrast, our method excludes these spurious points according to the distance to the deformed mesh, which is consistent with the fact that only points near the surface have positive effects in the rendering. We set the distance threshold $\varepsilon = 0.04m$ empirically in the experiments. As shown in both the comparison examples and sequential results of the supplementary video, our method can generate photo-realistic novel-view synthesis.

Fast fine-tuning. We demonstrate the effectiveness of our fast fine-tuning by comparing the predictions of the pre-trained networks (before fine-tuning) with results after fine-tuning. We can see in Table 3 that the prediction accuracy of the networks increases after fine-tuning. As shown in Figure 13 (e), the rendered textures before fine-tuning are blurry and lack of details. In contrast, the synthesized images after fine-tuning become more photo-realistic with more texture details (Figure 13 (c)). The fine-tuning on the four-view videos with 300 frames takes less than 1 hour, while training on the videos from scratch costs about 24 hours. This experiment demonstrates that the fast fine-tuning can improve the accuracy of both human performance capture and novel-view synthesis.

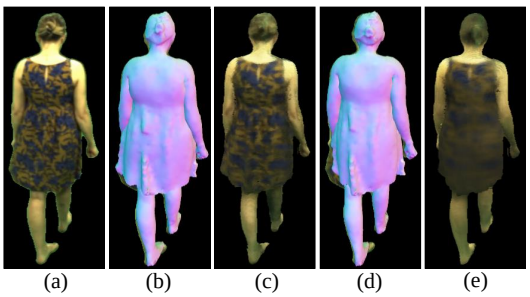


Fig. 13. Comparison between with and without fast fine-tuning. (a) The ground truth image. (b,d) The recovered geometry (overlayed on the image) with and without fast fine-tuning, respectively. (c,e) The novel-view synthesis result with and without fast fine-tuning, respectively.

Effectiveness of dynamic NeRF on performance capture. The silhouette constraint is adopted in the human performance capture methods [12], [13]. To testify effectiveness of dynamic NeRF on performance capture, we add the silhouette constraint in our framework and compare with our method. The quantitative results show that the accuracy even decreases with the silhouette constraint because the ambiguity in 2D silhouettes under complex human deformations may leads to erroneous alignment between the deformed template and observed points. In contrast, the dynamic NeRF forms dense constraints in both appearance and geometry to track the non-rigidly deforming surface, which gives rise to higher accuracy and stronger robustness in human performance capture. As also demonstrated in the comparison experiment of our method to [12], [13] (Sec. 4.3), our method can generate smoothly deforming geometry and appearance without using any temporal filter.

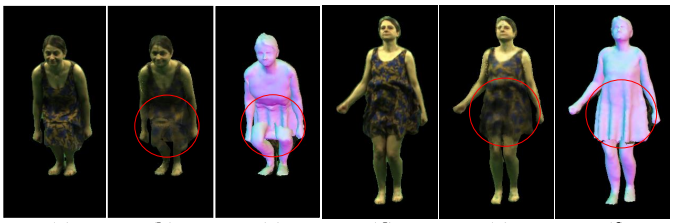
Fig. 14. Comparison to the human NeRF models in [20], [21]. (a) The ground truth image in a novel view. (b,c) The human NeRF model in NeuralActor [21] and HumanNeRF [20], respectively. (d) Our NeRF model. In each method, we show the novel-view synthesis and the reconstructed geometry (overlayed on the image). The local deformations of the moving dress (marked with circles) are not recovered accurately using the NeRF models in [20], [21], while our method reliably captures the clothing motion with the embedded graph prior.

Comparison to recent human NeRF models. To show the advantages of our dynamic NeRF, we compare it with the recent human NeRF models [20], [21]. Both NeuralActor [21] and HumanNeRF [20] model nonrigid human deformation with a residual displacement on top of the linear blend skinning so that they can represent the nonlinear local deformation that fails to be captured by the standard skinning adopted in [15], [16]. NeuralActor [21] learns the displacement

with a MLP from point positions and the body pose vector, while HumanNeRF [20] estimates it from the distances and directions of sample points relative to the skeleton joints and the point features. Since the codes of the two compared methods are not released, we implement their NeRF models and replace our NeRF model with them in our framework. The quantitative errors using the NeRF model of [20], [21] are listed in Table 3, and qualitative comparisons are shown in Fig. 14. Please refer to the supplementary video for the complete comparison of the sequence.

The human NeRF methods [20], [21] are mainly proposed for novel-view synthesis, while our method focuses on tracking the deforming template with dynamic human NeRF. From the comparison results, we can see the quality of synthesis images using the NeRF models of [20], [21] is comparable to us, but some local parts of the human are not reconstructed accurately in both the novel-view synthesis and geometry (e.g., the swing dress marked with circles in Fig. 14). Also, their recovered geometry is bumpy and noisy and has no explicit correspondences between different frames. The two compared methods do not use any motion prior to constrain the prediction of the displacement. Under complex and arbitrary deformations of loose clothes, the residual displacement fails to capture the clothing motion effectively in the under-constrained problem. In contrast, regularized by the explicit motion prior of the embedded graph, our method reliably captures the deformation of dressed humans, resulting in more accurate geometry tracking. In addition, our method obtains the space-time coherent correspondences across the videos by tracking the deforming template as [12], [13].

Limitations of NerfCap. Our method may fail to track the deforming geometry accurately under extreme cloth deformations (e.g., squatting with a skirt) that cannot be modeled by the embedded graph deformation. Some failure cases are shown in Fig. 15. The physics-based simulation method may be able to recover these extreme deformations of clothing, which we think is a promising future work. In addition, the accuracy may decrease when fitting on very long videos as shown in Table 3 and Fig. 10. The arbitrarily varying deformations of both the human body and clothing in long videos pose challenges on the learning of the canonical NeRF since using a single NeRF is hard to represent the tremendous variation space. This problem may be alleviated by learning temporally-varying NeRFs.



(a) (b) (c) (d) (e) (f)

Fig. 15. Failure examples of our method. (a,d) The ground truth image in a novel view. (b,e) The novel-view synthesis. (c,f) The recovered geometry (overlayed on the image). Our method may fail to recover extreme cloth deformations (surrounded by the circle).

4.6 Test on ZJU-MoCap data [15] and our data

To testify the effectiveness of our method on the data captured by general users, we first apply our method on ZJU-MoCap data [15]. ZJU-MoCap dataset provide 21-view synchronized videos, the ground-truth mask and SMPL model for each frame, and the parameters of all cameras. We randomly select a four-view video of 300 frames, and reconstruct a personalized template for the person. We train our network on the four-view video, and show the results of performance capture and novel-view synthesis in Figure 16 (first row). In addition, we capture our own data with a multi-camera system, and generate all the prerequisite data above using the methods described in the Sec. 3.1. The results of our method on our own data are shown in Figure 16 (second row). The results demonstrate the robustness and accuracy of our method on general data in terms of both human performance capture and novel-view synthesis.

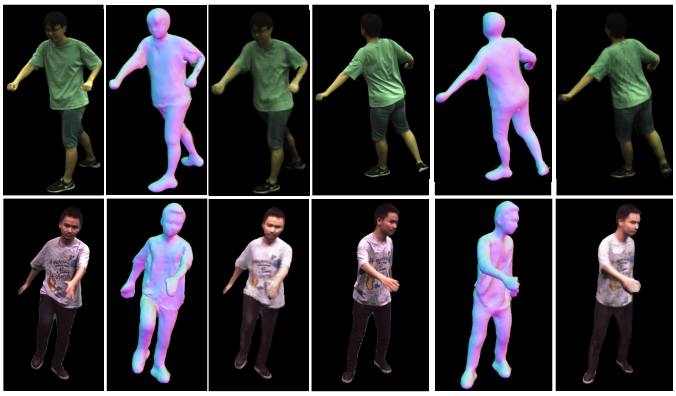


Fig. 16. Results of our method on ZJU-MoCap data [15] (top) and our data (bottom). For each example, we show the ground truth image in a novel view, the recovered geometry (overlayed on the image), and the novel-view synthesis. Note that the recovered geometry precisely overlays to the image and the novel-view synthesis is photo-realistic.

(a) (b) (c)

Fig. 17. Monocular human performance capture on the in-the-wild video [13] using our method. (a) The input image. (b) The recovered geometry (overlayed on the image). (c) The novel-view synthesis result.

4.7 Application: outdoor and monocular human performance capture

From a monocular video of a performer in outdoor scenes, capturing the human motion is very challenging due to complex backgrounds, serious occlusion and ambiguity

problem. Our method still can generate impressive results tested on a monocular and outdoor video. Figure 17 shows the results of our method on the in-the-wild video [13] captured by a single camera. Please refer to the supplementary video for results of the complete sequence. We can see that the recovered geometry robustly overlays to the input image and the novel-view synthesis is photo-realistic. This application can be adapted and used in many circumstances of virtual reality and augmented reality, e.g., telepresence and sports broadcasting. The realistic and immerse experience can be tremendously enhanced by recovering high-quality human geometry and appearance. For example, with a single camera, we can achieve a light-weight and realistic telepresence by showing 3D detailed humans during the remote communication. In the figure skating, we can observe the graceful dancing of the performer from free viewpoints. We believe high-quality human performance capture is valuable and promising with wide applications in virtual reality and augmented reality.

5 Conclusion

In this paper, we proposed a novel approach for human performance capture with dynamic neural radiance fields (NeRF). Our method jointly optimizes the human deformation and a canonical NeRF of the template mesh. We combine the non-rigid surface deformation with the skeleton-driven deformation to model moving humans in general clothing. Based on a deformation field produced from the human deformation, the canonical NeRF is mapped into an observation frame which is rendered into images. By comparing the rendered images with the observed images, the dynamic NeRF effectively captures the human geometry and appearance and simultaneously provides dense constraints to track the non-rigidly deforming surface. The experimental results on diverse datasets demonstrate the effectiveness of the proposed method.

Acknowledgment

This work was supported by the Natural Science Foundation of China under Grant Nos.62172364, U1713208, and Program for Changjiang Scholars, and the Fundamental Research Funds for the Central Universities (NJ2022028).

References

[1] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun, "Performance capture from sparse multi-view video," *ACM TOG*, vol. 27, no. 3, pp. 1–10, 2008.

[2] Y. Liu, C. Stoll, J. Gall, H.-P. Seidel, and C. Theobalt, "Markerless motion capture of interacting characters using multi-view image segmentation," in *CVPR*, 2011.

[3] D. Vlasic, P. Peers, I. Baran, P. Debevec, J. P. S. Rusinkiewicz, and W. Matusik, "Dynamic shape capture using multi-view photometric stereo," *ACM TOG*, vol. 28, no. 5, pp. 1–11, 2009.

[4] D. Vlasic, I. Baran, W. Matusik, and J. Popović, "Articulated mesh animation from multi-view silhouettes," *ACM TOG*, vol. 27, no. 3, pp. 1–9, 2008.

[5] G. Pons-Moll, S. Pujades, S. Hu, and M. J. Black, "ClothCap: Seamless 4D clothing capture and retargeting," *ACM TOG*, vol. 36, no. 4, pp. 73:1–73:15, 2017.

[6] M. Dou, S. Khamis, Y. Degtyarev, P. Davidson, S. Fanello, A. Kowdle, S. O. Escolano, C. Rhemann, D. Kim, J. Taylor, P. Kohli, V. Tankovich, and S. Izadi, "Fusion4D: Real-time performance capture of challenging scenes," *ACM TOG*, vol. 35, no. 4, pp. 1–13, Jul. 2016.

[7] R. Newcombe, D. Fox, and S. Seitz, "DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time," in *CVPR*, 2015.

[8] 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 *ICCV*, 2017.

[9] 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 *CVPR*, 2018.

[10] 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 *ECCV*, 2018, pp. 384–400.

[11] W. Xu, A. Chatterjee, M. Zollhöfer, H. Rhodin, D. Mehta, H.-P. Seidel, and C. Theobalt, "MonoPerfCap: Human performance capture from monocular video," *ACM TOG*, vol. 37, no. 2, pp. 1–15, Jul. 2018.

[12] M. Habermann, W. Xu, M. Zollhoefer, G. Pons-Moll, and C. Theobalt, "LiveCap: Real-time human performance capture from monocular video," *ACM TOG*, vol. 38, no. 2, pp. 14:1–14:17, 2019.

[13] ——, "DeepCap: Monocular human performance capture using weak supervision," in *CVPR*, 2020.

[14] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," in *ECCV*, 2020.

[15] S. Peng, Y. Zhang, Y. Xu, Q. Wang, Q. Shuai, H. Bao, and X. Zhou, "Neural Body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans," in *CVPR*, 2021.

[16] S. Peng, J. Dong, Q. Wang, S. Zhang, Q. Shuai, X. Zhou, and H. Bao, "Animatable neural radiance fields for modeling dynamic human bodies," in *ICCV*, 2021.

[17] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer, "D-NeRF: Neural radiance fields for dynamic scenes," in *CVPR*, 2021.

[18] E. Tretschk, A. Tewari, V. Golyanik, M. Zollhöfer, C. Lassner, and C. Theobalt, "Non-Rigid Neural Radiance Fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video," in *ICCV*, 2021.

[19] K. Park, U. Sinha, J. T. Barron, S. Bouaziz, D. B. Goldman, S. M. Seitz, and R. Martin-Brualla, "Nerfies: Deformable neural radiance fields," in *ICCV*, 2021.

[20] F. Zhao, W. Yang, J. Zhang, P. Lin, Y. Zhang, J. Yu, and L. Xu, "HumanNeRF: Generalizable neural human radiance field from sparse inputs," in *CVPR*, 2022.

[21] L. Liu, M. Habermann, V. Rudnev, K. Sarkar, J. Gu, and C. Theobalt, "Neural Actor: Neural free-view synthesis of human actors with pose control," *ACM TOG*, vol. 40, no. 6, pp. 1–16, 2021.

[22] M. Habermann, L. Liu, W. Xu, M. Zollhoefer, G. Pons-Moll, and C. Theobalt, "Real-time deep dynamic characters," *ACM TOG*, vol. 40, no. 4, pp. 1–16, August 2021.

[23] Q. Ma, S. Saito, J. Yang, S. Tang, and M. J. Black, "SCALE: Modeling clothed humans with a surface codec of articulated local elements," in *CVPR*, 2021.

[24] S. Saito, J. Yang, Q. Ma, and M. J. Black, "SCANimate: Weakly supervised learning of skinned clothed avatar networks," in *CVPR*, 2021.

[25] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, "PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization," in *ICCV*, 2019.

[26] S. Saito, T. Simon, J. Saragih, and H. Joo, "PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization," in *CVPR*, 2020.

[27] H. Zhu, X. Zuo, S. Wang, X. Cao, and R. Yang, "Detailed human shape estimation from a single image by hierarchical mesh deformation," in *CVPR*, 2019.

[28] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu, "DeepHuman: 3D human reconstruction from a single image," in *ICCV*, 2019.

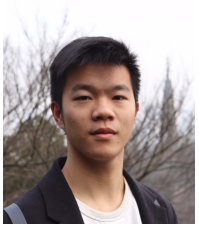
[29] T. Alldieck, G. Pons-Moll, C. Theobalt, and M. Magnor, "Tex2Shape: Detailed full human body geometry from a single image," in *ICCV*, 2019.

[30] H. Zhao, J. Zhang, Y.-K. Lai, Z. Zheng, Y. Xie, Y. Liu, and K. Li, "High-fidelity human avatars from a single rgb camera," in *CVPR*, 2022.

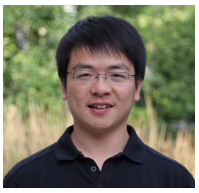
[31] B. Jiang, Y. Hong, H. Bao, and J. Zhang, "SelfRecon: Self reconstruction your digital avatar from monocular video," in *CVPR*, 2022.

[32] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM TOG*, vol. 34, no. 6, pp. 248:1–248:16, 2015.
[33] 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 *CVPR*, 2017, pp. 6050–6059.
[34] 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, pp. 459–468.
[35] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, "End-to-end recovery of human shape and pose," in *CVPR*, 2018, pp. 7122–7131.
[36] 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 *ECCV*, 2016, pp. 561–578.
[37] N. Kolotouros, G. Pavlakos, and K. Daniilidis, "Convolutional mesh regression for single-image human shape reconstruction," in *CVPR*, 2019.
[38] N. Kolotouros, G. Pavlakos, M. J. Black, and K. Daniilidis, "Learning to reconstruct 3D human pose and shape via model-fitting in the loop," in *ICCV*, 2019.
[39] H. Jiang, J. Cai, and J. Zheng, "Skeleton-aware 3D human shape reconstruction from point clouds," in *ICCV*, 2019.
[40] S. Prokudin, C. Lassner, and J. Romero, "Efficient learning on point clouds with basis point sets," in *ICCV*, 2019.
[41] K. Wang, J. Xie, G. Zhang, L. Liu, and J. Yang, "Sequential 3D human pose and shape estimation from point clouds," in *CVPR*, 2020, pp. 7275–7284.
[42] A. Kanazawa, J. Y. Zhang, P. Felsen, and J. Malik, "Learning 3D human dynamics from video," in *CVPR*, 2019.
[43] J. Y. Zhang, P. Felsen, A. Kanazawa, P. Felsen, and J. Malik, "Predicting 3D human dynamics from video," in *ICCV*, 2019.
[44] X. Chen, A. Pang, W. Yang, Y. Ma, L. Xu, and J. Yu, "SportsCap: Monocular 3D human motion capture and fine-grained understanding in challenging sports videos," *IJCV*, pp. 2846–2864, 2021.
[45] A. Arnab, C. Doersch, and A. Zisserman, "Exploiting temporal context for 3D human pose estimation in the wild," in *CVPR*, 2019.
[46] 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 *CVPR*, 2019.
[47] B. L. Bhatnagar, G. Tiwari, C. Theobalt, and G. Pons-Moll, "Multi-Garment Net: Learning to dress 3D people from images," in *ICCV*, 2019.
[48] Q. Ma, J. Yang, A. Ranjan, S. Pujades, G. Pons-Moll, S. Tang, and M. J. Black, "Learning to dress 3D people in generative clothing," in *CVPR*, 2020.
[49] C. Patel, Z. Liao, and G. Pons-Moll, "TailorNet: Predicting clothing in 3D as a function of human pose, shape and garment style," in *CVPR*, 2020.
[50] B. Jiang, J. Zhang, Y. Hong, J. Luo, L. Liu, and H. Bao, "BCNet: Learning body and cloth shape from a single image," in *ECCV*, 2020.
[51] 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.
[52] J. P. Lewis, M. Cordner, and N. Fong, "Pose space deformation: A unified approach to shape interpolation and skeleton-driven deformation," in *SIGGRAPH*, 2000.
[53] S. Peng, Z. Xu, J. Dong, Q. Wang, S. Zhang, Q. Shuai, H. Bao, and X. Zhou, "Animatable implicit neural representations for creating realistic avatars from videos," *arXiv:2105.02872*, 2022.
[54] H. Xu, T. Alldieck, and C. Sminchisescu, "H-NeRF: Neural radiance fields for rendering and temporal reconstruction of humans in motion," in *Proceedings of International Conference on Neural Information Processing Systems*, 2021.
[55] K. Gong, X. Liang, Y. Li, Y. Chen, M. Yang, and L. Lin, "Instance-level human parsing via part grouping network," in *ECCV*, 2018.
[56] H. Joo, T. Simon, and Y. Sheikh, "Total capture: A 3D deformation model for tracking faces, hands, and bodies," in *CVPR*, 2018.
[57] R. W. Sumner, J. Schmid, and M. Pauly, "Embedded deformation for shape manipulation," *ACM TOG*, vol. 26, no. 3, 2007.
[58] M. Oechsle, S. Peng, and A. Geiger, "UNISURF: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction," in *ICCV*, 2021.
[59] Q. Wang, Z. Wang, K. Genova, P. Srinivasan, H. Zhou, J. T. Barron, R. Martin-Brualla, N. Snavely, and T. Funkhouser, "IBRNet: Learning multi-view image-based rendering," in *CVPR*, 2021.
[60] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference for Learning Representations*, 2015.

Kangkan Wang is currently an associate professor in the School of Computer Science and Engineering of Nanjing University of Science and Technology. Before that, he was an assistant professor at the Institute of Computing Technology, Chinese Academy of Sciences. He received the BS degree in computer science from Northwestern Polytechnical University in 2009 and the PhD degree from the State Key Lab of CAD&CG of Zhejiang University in 2015. His research interests include human body reconstruction, 3D reconstruction, non-rigid reconstruction and motion tracking.



Sida Peng is currently a Ph.D. student in computer science at Zhejiang University, advised by Dr. Xiaowei Zhou. He received B.E. degree in information engineering from Zhejiang University in 2018. His research interests include 3D reconstruction and object pose estimation.



Xiaowei Zhou is a Research Professor of Computer Science at Zhejiang University, China. He obtained his PhD degree from The Hong Kong University and Science and Technology, after which he was a postdoctoral researcher at the GRASP Lab, University of Pennsylvania. His research interests include 3D reconstruction, understanding and synthesis of objects, humans and scenes, with applications in VR/AR and robotics. He is on the editorial board of IJCV, served as area chairs of CVPR'21 and ICCV'21, and co-organized the series of Geometry Meets Deep Learning Workshops (GMDL).

Jian Yang is a Chang-Jiang professor in the School of Computer Science and Engineering of Nanjing University of Science and Technology (NUST). He received the PhD degree from NUST in 2002. In 2003, he was a Postdoctoral researcher at the University of Zaragoza. From 2004 to 2006, he was a Postdoctoral Fellow at Biometrics Centre of Hong Kong Polytechnic University. From 2006 to 2007, he was a Postdoctoral Fellow at Department of Computer Science of New Jersey Institute of Technology. He is the author of more than 200 scientific papers in pattern recognition and computer vision. His papers have been cited more than 6000 times in the Web of Science, and 15000 times in the Scholar Google. His research interests include pattern recognition, computer vision and machine learning. He is a Fellow of IAPR.

Guofeng Zhang now is a Professor at State Key Lab of CAD&CG, Zhejiang University. He received his BS and Ph.D. degrees in Computer Science from Zhejiang University, in 2003 and 2009, respectively. He received the National Excellent Doctoral Dissertation Award, the Excellent Doctoral Dissertation Award of China Computer Federation and the best paper award of ISMAR 2020. His research interests include structure-from-motion, SLAM, 3D reconstruction, augmented reality, video segmentation, and editing.