Animatable Implicit Neural Representations for Creating Realistic Avatars From Videos

Xiaowei Zhou, Sida Peng, Zhen Xu, Graduate Student Member, IEEE, Junting Dong, Qianqian Wang, Shangzhan Zhang, Graduate Student Member, IEEE, Qing Shuai, and Hujun Bao, Member, IEEE

Abstract—This paper addresses the challenge of reconstructing an animatable human model from a multi-view video. Some recent works have proposed to decompose a non-rigidly deforming scene into a canonical neural radiance field and a set of deformation fields that map observation-space points to the canonical space, thereby enabling them to learn the dynamic scene from images. However, they represent the deformation field as translational vector field or SE(3) field, which makes the optimization highly under-constrained. Moreover, these representations cannot be explicitly controlled by input motions. Instead, we introduce blend weight fields to produce the deformation fields. Based on the skeleton-driven deformation, blend weight fields are used with 3D human skeletons to generate observation-to-canonical and canonical-to-observation correspondences. Since 3D human skeletons are more observable, they can regularize the learning of deformation fields. Moreover, the blend weight fields can be combined with input skeletal motions to generate new deformation fields to animate the human model. To improve the quality of human modeling, we further represent the human geometry as a signed distance field in the canonical space. Additionally, a neural point displacement field is introduced to enhance the capability of the blend weight field on modeling detailed human motions. Experiments show that our approach significantly outperforms recent human modeling methods.

Index Terms—Human modeling, implicit neural representations, view synthesis.

I. INTRODUCTION

ENDERING animatable human characters has many applications such as free-viewpoint videos, telepresence, video games and movies. The core step is to reconstruct animatable human models, which tends to be expensive and time-consuming in traditional pipelines due to two factors. First, human reconstruction generally relies on complicated hardware, such as a dense array of cameras [1], [2] or depth sensors [3], [4]. Second, human animation requires skilled artists to manually create a skeleton suitable for the human model and carefully design skinning weights [5] to achieve realistic animation, which takes countless human labor.

In this work, we aim to reduce the cost of human reconstruction and animation, to enable the creation of digital humans at scale. Specifically, we focus on the problem of reconstructing animatable humans from multi-view videos, as illustrated in Fig. 1. However, this problem is extremely challenging. There are two core questions we need to answer: how to represent animatable human models and how to learn this representation from videos?

Recently, neural radiance fields (NeRF) [6] has proposed a representation that can be efficiently learned from images with a differentiable renderer. It represents static 3D scenes as color and density fields, which work particularly well with volume rendering techniques. To extend NeRF to handle non-rigidly deforming scenes, [7], [8] decompose a video into a canonical NeRF and a set of deformation fields that transform observation-space points at each video frame to the canonical space. The deformation field is represented as translational vector field [8] or SE(3) field [7]. Although they can handle some dynamic scenes, they are not suited for representing animatable human models due to two reasons. First, jointly optimizing NeRF with translational vector fields or SE(3) fields without motion prior is an extremely under-constrained problem [8], [9]. Second, they cannot synthesize novel scenes given input motions.

To overcome these problems, we propose a novel motion representation named blend weight field. Based on the skeleton-driven deformation framework [5], blend weight fields are combined with 3D human skeletons to generate deformation fields. This representation has two advantages. First, since the human skeleton is easy to track [10], it does not need to be jointly optimized and thus provides an effective regularization on the learning of deformation fields. Second, we are able to explicitly animate the canonical human model with input human skeletons, in contrast to previous methods [8], [9].

To learn the human model from videos, we first leverage blend weight fields to produce deformation fields that transform 3D points from the observation space to the canonical space, and then render the canonical NeRF into video frames. The model parameters are optimized to minimize the photometric loss between rendered frames and input frames. In practice, experiments show that it is difficult to learn blend weight fields...
from scratch. Therefore, we propose to obtain blend weight fields from SMPL models [11] and then combine it with a neural residual field to better model human motions.

Based on the proposed human representation, we add two extensions to improve the performance of human modeling. Considering that NeRF tends to produce noisy geometry surface, the human geometry is represented as a signed distance field, which has a well-defined surface at the zero-level set and thus facilitates more direct regularization on the geometry learning. However, we found that jointly optimizing the signed distance field with neural residual vector fields tends to get into local minima. To solve this issue, a neural point displacement field is introduced to model motion details, instead of using neural residual vector field. Experiments demonstrate that neural point displacement field works well with the signed distance field.

We evaluate our approach on the Human3.6M [12], ZJU-MoCap [13] and MonoCap [14, 15, 16] datasets. Experiments show that our approach achieve state-of-the-art performance on image synthesis. In addition, we demonstrate that using signed distance field significantly improve the reconstruction performance on SyntheticHuman [16].

In the light of previous work, this work has the following contributions: i) We reconstruct animatable human models from videos using a novel framework, which represents a dynamic human with a canonical neural field and a pose-driven deformation field. ii) We compares different implicit neural representations for human modeling from videos. iii) Our approach demonstrates significant performance improvement on image rendering and 3D reconstruction compared to recent methods on the Human3.6M, ZJU-MoCap, MonoCap and SyntheticHuman datasets.

A preliminary version of this work appeared in ICCV 2021 [17]. Here, the work is extended in the following ways. First, the signed distance field is used to enhance the quality of human geometry. We validate the effectiveness of signed distance field on 3D reconstruction on a synthetic dataset. Second, we introduce the neural point displacement field to model accurate human motions in conjunction with the blend weight field. Third, we compare with more recent methods and additionally perform experiments on a monocular dataset [14, 15]. More ablation studies are added to evaluate the proposed components. The code has been publicly available at https://github.com/zju3dv/animatable_nerf, which has received more than 400 stars.

**II. RELATED WORK**

**Human reconstruction:** Modeling human characters is the first step of traditional animation pipelines. To achieve high-quality reconstruction, most methods rely on complicated hardware [2], [3], [4], [18], [19]. Recently, some works [6], [20], [21], [22] have attempted to learn 3D representations from images with differentiable renderers, which reduces the number of input camera views and achieves impressive reconstruction results. However, they have difficulty in recovering reasonable 3D human shapes when the camera views are too sparse, as shown in [13]. Instead of optimizing the network parameters per scene, [23], [24], [25], [26] utilize networks to learn human shape priors from ground-truth 3D data, allowing them to reconstruct human shapes from even a single image.

**Human animation:** Skeletal animation [5], [27] is a common approach to animate human models. It first creates a scale-appropriate skeleton for the human mesh and then assigns each mesh vertex a blend weight that describes how the vertex position deforms with the skeleton. Skinned multi-person linear model (SMPL) [11] learns a skeleton regressor and blend weights from a large amount of ground-truth 3D meshes. Based on SMPL, some works [28], [29], [30], [31], [32] reconstruct an animated human mesh from sparse camera views. However, SMPL only describes the naked human body and thus cannot be directly used to render photorealistic images. To overcome this problem, [33], [34], [35] apply vertex displacements to the SMPL model to capture the human clothing and hair. [36] proposes a 2D warping method to deform the SMPL model to fit the input image. Recent implicit function-based methods [37], [38], [39] have exhibited state-of-the-art reconstruction quality. [40], [41] combine implicit function learning with the SMPL model to obtain detailed animatable human models. [42] combines a set of local implicit functions with human skeletons to represent dynamic humans. [43], [44], [45], [46], [47] propose to animate implicit neural representations with the linear blend skinning algorithm.

**Neural rendering:** To reduce the requirement for the reconstruction quality, some methods [48], [49], [50], [51], [52] improve the rendering pipeline with neural networks. Based on the advances in image-to-image translation techniques [53], [54], [55], [56] train a network to map 2D skeleton images to target rendering results. Although these methods can synthesize photorealistic images under novel human poses, they have difficulty...
in rendering novel views. To improve the performance of novel view synthesis, [48], [49], [51], [57], [58], [59], [60] introduce 3D representations into the rendering pipeline. [49] establishes neural texture maps and uses UV maps to obtain feature maps in the image space, which is then interpreted into images with a neural renderer. [51], [57] reconstruct a point cloud from input images and learn a 3D feature for each point. Then, they project 3D features into a 2D feature map and employ a network to render images. However, 2D convolutional networks have difficulty in rendering inter-view consistent images, as shown in [20].

To solve this problem, [6], [21], [61], [62], [63], [64] interpret features into colors in 3D space and then accumulate them into 2D images. In the field of human modeling, [13], [65], [66], [67], [68] represent 3D human models as implicit neural representations and optimize network parameters from images with differentiable volume rendering. [13] combines neural radiance field with the SMPL model, allowing it to handle dynamic humans and synthesize photorealistic novel views from very sparse camera views. Reconstructing 3D humans from videos is a fast growing field, and there are many concurrent works [69], [70], [71], [72], [73], [74], [75], [76]. Similar to [17], [71], [72], [74], [76] leverage the LBS model to establish observation-to-canonical correspondences, which enables them to aggregate temporal observations in the input video.

III. Method

Given a multi-view video of a performer, our task is to reconstruct an animatable human model that can be used to synthesize free-viewpoint videos of the performer under novel human poses. The cameras are synchronized and calibrated. For each frame, we assume the 3D human skeleton is given, which can be obtained with marker-based or marker-less pose estimation systems [10], [12]. For each image, [77] is used to extract the foreground human mask, and the values of the background image pixels are set as zero.

The overview of our approach is shown in Fig. 2. We decompose a non-rigidly deforming human body into a canonical human model represented by a neural radiance field (Section III-A) and a per-frame blend weight field (Section III-B) that is used to establish correspondences between the observation space and canonical space. Section III-C introduces two extensions to the human representation to enhance the quality of human modeling. Then we discuss how to learn the representation on the multi-view video (Section III-D). Based on blend weight fields, we are able to animate the canonical human model (Section III-E).

A. Representing Videos With Neural Radiance Fields

NeRF represents a static scene as a continuous volumetric representation. For any 3D point, it takes a spatial position \( x \) and viewing direction \( d \) as input to a neural network and outputs a volume density \( \sigma \) and color \( c \).

Inspired by [7], [8], we extend NeRF to represent the dynamic human body by introducing deformation fields, as shown in Fig. 2. Specifically, for each video frame \( i \in \{1, \ldots, N\} \), we define a deformation field \( T_i \) that transforms observation-space points to the canonical space. Given the canonical-frame density model \( F_\sigma \), the density model at frame \( i \) can be thus defined as:

\[
\sigma_i(x, z_i(x)) = F_\sigma(\gamma_\times(T_i(x))),
\]

where \( z_i(x) \) is the shape feature in the original NeRF, and \( \gamma_\times \) is the positional encoding [6] for spatial location.

When predicting the color, we define a per-frame latent code \( \ell_i \) to encode the state of the human appearance in frame \( i \). Similarly, with the canonical-frame color model \( F_c \), the color model at frame \( i \) can be defined as:

\[
c_i(x) = F_c(z_i(x), \gamma_d(d), \ell_i),
\]

where \( \gamma_d \) is the positional encoding for viewing direction.

There are several ways to represent the deformation field, such as translational vector field [8], [9] and SE(3) field [7]. However, as discussed in [7], [9], optimizing a radiance field together with a deformation field is an ill-posed problem. To solve this issue, [7], [9] propose many regularization techniques to facilitate the training, which makes the optimization process complex. Moreover, their representations cannot robustly generate new deformation fields given novel motion sequences.
B. Blend Weight Fields

Considering that we aim to model dynamic humans, it is natural to leverage the human priors to learn the deformation field, which helps us to solve the under-constrained problem. Specifically, we construct the deformation field based on the 3D human skeleton and the skeleton-driven deformation framework [5].

The human skeleton defines $K$ parts, which produce $K$ transformation matrices $\{G_k\} \in SE(3)$. The detailed derivation is listed in the supplementary material, available online. In the linear blend skinning algorithm [5], a canonical-space point $v$ is transformed to the observation space using

$$v' = \left( \sum_{k=1}^{K} w(v)_k G_k \right) v,$$

where $w(v)_k$ is the blend weight of $k$th part. Similarly, for an observation-space point $x$, if we know its corresponding blend weights, we are able to transform it to the canonical space using

$$x' = \left( \sum_{k=1}^{K} w^o(x)_k G_k \right)^{-1} x,$$

where $w^o(x)$ is the blend weight function defined in the observation space. To obtain the blend weight field, a natural idea is to define a function that maps a 3D point to blend weights, which then gives the dynamic radiance fields based on (1), (2) and (4). However, experiments show that learning the blend weight field from scratch is still ill-posed and is prone to local minima.

Neural residual vector field: To solve this problem, we propose to leverage the human priors in 3D statistical body models [11], [78], [79], [80]. Specifically, for any 3D point, we assign an initial blend weight based on the body model and then use a network to learn a residual vector, resulting in the subject-specific blend weight field. The residual vector fields for all training video frames are implemented using a single MLP network $F_{\Delta w} : (x, S_i) \rightarrow \Delta w$, where $\psi_i$ is a per-frame learned latent code and $\Delta w_i$ is a vector $\in \mathbb{R}^K$. The resulted blend weight field at frame $i$ is defined as:

$$w_i(x) = \text{norm}(F_{\Delta w}(x, \psi_i) + w^o(x, S_i)),$$

where $w^o$ is the initial blend weights that are computed based on the statistical body model $S_i$, and we define $\text{norm}(w) = w / \sum w_i$. Without loss of generality, we adopt SMPL [11] as the body model, which can be obtained by fitting the SMPL model to the 3D human skeleton [10]. Note that this idea can also apply to other human models [78], [79], [80]. To compute $w^o$, we take the strategy proposed in [40], [81]. For any 3D point, we first find the closest surface point on the SMPL mesh. Then, the target blend weight is computed by performing barycentric interpolation of blend weights of three vertices on the corresponding mesh facet.

To animate the template NeRF, we additionally define a blend weight field $w_{\text{can}}$ at the canonical space. The SMPL blend weight field $w^o$ is calculated using the canonical SMPL model, and the residual vector field $F_{\Delta w}$ is conditioned on an additional latent code $\psi^\text{can}$. We utilize the consistency between blend weights to optimize the residual vector field $F_{\Delta w}$, which is described in Section III-D. During animation, the canonical blend weight field is used to compute blend weights for observation-space coordinates under unseen human poses, which is described in Section III-E.

Instead of calculating blend weights of novel human poses from the canonical blend weight field $w_{\text{can}}$, an alternative method is to define the pose-dependent residual vector field $F'_{\Delta w} : (x, S) \rightarrow \Delta w$, which enables us to directly predict the blend weight field under human pose $S$ using (5). However, the input coordinate $x$ for $F'_{\Delta w}$ is in the observation space, whose value could vary significantly with human pose, making the residual vector field $F'_{\Delta w}$ difficult to generalize to unseen poses.

C. Improving the Learning of Human Geometry

Although NeRF has shown impressive performance on view synthesis, it does not explicitly constrain the learned geometry and tends to produce the noisy surface. To enhance the reconstruction quality, we model the human geometry using the signed distance field. In contrast to density field, signed distance field has a well-defined surface at the zero-level set, which facilitates more direct regularization on the geometry learning and generally achieves better reconstruction performance [82], [83]. In practice, since signed distance is a scalar, we can predict the signed distance $s$ for each 3D point $x$ using the same geometry network $F_g : x \rightarrow s$. We apply Eikonal constraint [84] to enforce the network prediction to conform with the property of signed distance field, which will be described in Section III-D. However, we empirically find that it is difficult to jointly learn the signed distance field with the neural residual vector field from videos.

Neural point displacement field: To overcome this challenge, we propose to model human motions by combining the SMPL blend weight field with a neural point displacement field, instead of using the neural residual vector field. Specifically, for an observation-space point $x$ at frame $i$, we first obtain its blend weights from the SMPL model and warp it to the canonical space using (4), resulting in the transformed point $x'$. Then, we use a displacement field to deform the point $x'$ to the surface. Denote the displacement field as $F_{\Delta x} : (x, S_i) \rightarrow \Delta x_i$, where $S_i$ is the 3D human pose at frame $i$. The final point is $x' + F_{\Delta x}(x', S_i)$. The displacement field is implemented as an MLP network.

Experiments show that the point displacement field $F_{\Delta x}$ can generalize to unseen human poses. A plausible reason is that $F_{\Delta x}$ takes the canonical-space coordinate as input, whose value is similar across different human poses. Compared with neural residual vector field, the point displacement field is more convenient to use, as neural residual vector field additionally requires us to calculate blend weights under novel poses during animation. More details can be found in Section III-E.

D. Training

Based on the proposed canonical human representations and motion representations, we can construct two types of animatable human representations: (1) NeRF with residual vector field
(NeRF-RVF). (2) SDF with point displacement field (SDF-PDF). To verify the advantage of point displacement field over neural residual vector field, we additionally test a type of animatable human representation: NeRF with point displacement field (NeRF-PDF). We describe the training of these three models in this section.

Given an animatable human representation, we can use volume rendering techniques [6], [85] to synthesize images of particular viewpoints for each video frame $i$. The near and far bounds of volume rendering are estimated by computing the 3D boxes that bound the SMPL meshes. When the density field is used to model the human geometry, we directly use the rendering strategy in [6] to render pixel colors. The model parameters are optimized over the video by minimizing the difference between the rendered pixel color $C_i(r)$ and the observed pixel color $C_i'(r)$:

$$L_{rgb} = \sum_{r \in R} \| \bar{C}_i(r) - C_i(r) \|_2,$$

where $R$ is the set of rays passing through image pixels.

When the neural residual vector field is used to represent the human motion, we introduce a consistency loss to learn the weight field $w^{can}$ at the canonical space. As shown by (3) and (4), two corresponding points at canonical and observation spaces should have the same blend weights. For an observation-space point $x$ at frame $i$, we map it to the canonical-space point $T_i(x)$ using (4). The consistency loss is defined as:

$$L_{nsf} = \sum_{x \in X} \| w_i(x) - w^{can}(T_i(x)) \|_1,$$

where $\lambda_i^c$ is the set of 3D points sampled within the 3D human bounding box at frame $i$. For the model “NeRF-RVF”, we use the rendering loss $L_{rgb}$ and consistency loss $L_{nsf}$ for training, which is defined as:

$$L_{NeRF-RVF} = L_{rgb} + L_{nsf}. \quad (8)$$

**Learning of model “SDF-PDF”:** To optimize the SDF from the video, we first convert predicted signed distances into densities using the strategy in VolSDF [83] and then synthesize pixel colors through the volume rendering. The mask loss and the Eikonal term [84] are used for learning the SDF. To supervise the SDF with the mask, we find the minimal SDF value $s_i^m$ of sampled points along the ray $r$ and apply the binary cross entropy loss BCE:

$$L_{mask} = \sum_{r \in R} \text{BCE}(\text{sigmoid}(-\rho s_i^m), M_i(r)), \quad (9)$$

where $M_i(r) \in \{0, 1\}$ is the ground-truth mask value. Similar to [22], we set $\rho$ as 100 and multiply it by 2 every 10000 iterations. The number of multiplications is up to 5. We sample a set of points $X_i$ in the observation space and apply the Eikonal term on these sampled points:

$$L_E = \sum_{x \in X_i} (\| \nabla F_g(T(x, S_i)) \|_2 - 1)^2. \quad (10)$$

When the neural point displacement field is used to produce the deformation field, we apply the regularization to predicted displacements, which is defined as:

$$L_{\Delta x} = \sum_{x \in X'} \| F_{\Delta x}(x, S_i) \|_2^2, \quad (11)$$

where $X'$ is the set of 3D points sampled in the canonical space. The final loss for training model “SDF-PDF” is

$$L_{SDF-PDF} = L_{rgb} + L_{mask} + 0.01L_E + 0.01L_{\Delta x}. \quad (12)$$

**Learning of model “NeRF-PDF”:** When the canonical human model is represented as NeRF and the human motion is predicted based on neural point displacement field, we use the rendering loss $L_{rgb}$ and displacement regularization term $L_{\Delta x}$ for training, which is defined as:

$$L_{NeRF-PDF} = L_{rgb} + 0.01L_{\Delta x}. \quad (13)$$

**E. Animation**

After training, we can use the deformation field to animate the canonical human model. Given the novel human pose $S\text{new}$, the deformation field warps observation-space points to the canonical space, which are then fed into the geometry and color models. Note that the color model is conditioned on a latent code, which is not defined for unseen human pose. Therefore, we select the latent code of the training human pose that is nearest to the novel human pose $S\text{new}$.

When the deformation field is modeled using the neural residual vector field, we need to additionally optimize blend weights under novel poses based on the canonical blend weight field. Specifically, for the novel human pose $S\text{new}$, our method first computes the SMPL blend weight field $w^c$. Then, the blend weight field $w^\text{new}$ for the novel human pose is defined as:

$$w^\text{new}(x, \psi^\text{new}) = \text{norm}(F_{\Delta w}(x, \psi^\text{new}) + w^c(x, S\text{new})), \quad (14)$$

where the $F_{\Delta w}$ is conditioned on a new latent code $\psi^\text{new}$. Based on the $w^\text{new}$ and (4), we can generate the deformation field $T\text{new}$ for the novel human pose. The parameters of $\psi^\text{new}$ are optimized using

$$L_{\text{new}} = \sum_{x \in X'} \| w^\text{new}(x) - w^\text{can}(T\text{new}(x)) \|_1, \quad (15)$$

where $X'$ is the set of 3D points sampled within the human box under the novel human pose. Note that we fix the parameters of $w^\text{can}$ during training. In practice, we train neural residual vector fields under multiple novel human poses simultaneously. This is implemented by conditioning $F_{\Delta w}$ on multiple latent codes. With the deformation field $T\text{new}$, our method uses (1) and (2) to produce the human model under the novel human pose.

When the neural point displacement field $F_{\Delta x}$ is used, we can establish the observation-to-canonical correspondences by first warping observation-space points to the canonical space and then deforming them using $F_{\Delta x}$. In contrast to neural blend weight field, the pose-dependent displacement field does not require the additional optimization during animation, which is easier to use.
IV. IMPLEMENTATION DETAILS

The network of NeRF-based human model “NeRF-RVF” adopts the same architecture as the original NeRF [6], while the network of SDF-based human model “SDF-PDF” closely follows the original IDR [22]. “NeRF-PDF” also adopts the network of the original IDR [22]. The networks of $F_{z,w}$ and $F_{z,x}$ consist of nine fully connected layers. More details of network architectures are described in the supplementary material, available online. The appearance code $\ell_i$ and residual vector field code $\psi_i$ both have dimensions of 128.

Training: The Adam optimizer [86] is adopted for the training. The learning rate starts from $5e^{-3}$ and decays exponentially to $5e^{-5}$ along the optimization. The training is conducted on a 2080 Ti GPU. For a three-view video of 300 frames, the training takes around 200 k iterations to converge. During animation, we use the same optimizer and learning rate scheduler to optimize the neural blend weight field. For 200 novel human poses, the optimization takes around 10 k iterations to converge. More details about training can be found in the supplementary material, available online.

V. EXPERIMENTS

In experiments, we evaluate three types of animatable human representations, including “NeRF-RVF”, “SDF-PDF”, and “NeRF-PDF”, which are described in Section III-D. The model “NeRF-RVF” is our conference version [17]. To fairly compare “NeRF-RVF” and “SDF-PDF”, we update the network of “NeRF-RVF” to the same architecture as “SDF-PDF” and denote it as “NeRF-RVF*”.

A. Dataset and Metrics

Human3.6M [12] records multi-view videos with 4 cameras and collects human poses using the marker-based motion capture system. It includes multiple human subjects performing complex actions. We select representative human actions, split the videos into training and test frames, and perform experiments on subjects S1, S5, S6, S7, S8, S9, and S11. Three cameras are used for training and the remaining camera is selected for test. ZJU-MoCap [13] records 9 multi-view videos with 21 cameras and collects human poses using the marker-less motion capture system. Following the experimental protocol in [13], we select four uniformly distributed cameras as training input and test on the remaining cameras. MonoCap consists of two videos from DeepCap dataset [14] and two videos from DynaCap dataset [15]. We use one camera view for training and select ten uniformly distributed cameras for test. We select a clip of each video to perform experiments. Each clip has 300 frames for training and 300 frames for evaluating novel pose synthesis.

SyntheticHuman is a synthetic dataset created by [16], which contains 7 animated 3D characters. 4 human characters perform rotation while holding A-pose, which are rendered into monocular videos. Another 3 human characters perform random actions, which are rendered with four cameras. All video frames and camera views are used for training. This dataset is only used to evaluate the performance on 3D reconstruction.

Metrics: For 3D reconstruction, we follow [24] to use two metrics: point-to-surface euclidean distance (P2S) and Chamfer distance (CD). Units for the two metrics are in cm. For image synthesis, we follow [6] to evaluate our method using two metrics: peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM).

B. Performance on Image Synthesis

Our method is compared with previous methods that train a separate network for each video. We do not compare with [66], [88], because they do not release the source code. Detailed descriptions of the baseline methods are presented in the supplementary material, available online.

Results on the Human3.6M dataset: Table I compares our method with [8], [13], [49], [51], [87] on image synthesis. The results indicate that our models with the IDR network outperforms the baseline methods. Comparing model “NeRF-RVF*” with model “NeRF-RVF” shows that the IDR network largely improves the performance of “NeRF-RVF”, which will be further discussed in Section VI.

In Figs. 3 and 4, we present qualitative results of our method and baseline methods on novel view synthesis of training and novel poses. Our method produces photo-realistic rendering results and outperforms baseline methods. We can see that [51] has difficulty in controlling the rendering viewpoint and tend to synthesize contents of training views. As shown in the third person of Fig. 4, they render the human back that is seen during training. Although [13] synthesizes high-quality images on training poses, it struggles to give reasonable rendering results on novel human poses. In contrast, our method explicitly animates the canonical human model using a pose-driven deformation field, which is similar to the classical graphics pipeline and has better controllability on the image generation process than CNN-based methods. The supplementary material presents more qualitative results, available online.
Fig. 3. Novel view synthesis under training human poses. The first and second rows present results from the ZJU-MoCap dataset and the Human3.6 M dataset, respectively. “NeRF-RVF∗” means “NeRF-RVF” with the IDR network.

Fig. 4. Qualitative results of novel pose synthesis. Figures on the first row are results from the ZJU-MoCap dataset. The second row comes from the MonoCap dataset. The last two rows are from the Human3.6 M dataset. “NeRF-RVF∗” indicates “NeRF-RVF” with the IDR network.

Results on the ZJU-MoCap dataset: The quantitative results are listed in Table II. When trained on 4 camera views, our method is competitive to [13] on novel view synthesis of training poses and outperforms [13] on novel pose synthesis. Figs. 3 and 4 present qualitative results of models that are trained on 4 camera views. When trained on the single camera view, our method significantly outperforms [13] on both training and novel poses.

Results on the MonoCap dataset: Table IV summarizes the quantitative comparison of novel view synthesis between our methods and [87], [13], [51] under training and novel human
pose. Specifically, our methods all perform better than the baselines on novel view synthesis under novel human poses. And the pose-dependent deformation field methods NeRF-PDF and SDF-PDF outperform [13] on monocular novel view synthesis. Fig. 4 presents the qualitative comparison of image synthesis of our methods and baselines on MonoCap dataset, which show that our model is able to produce photo-realistic novel view synthesis under novel human poses even with only one input view.

C. Performance on 3D Reconstruction

To validate our method on 3D reconstruction, we compare with [8], [13], [87]. Because [51] does not reconstruct the human geometry and [66], [88] do not release the code, we do not compare with them. We use the Marching Cubes algorithm [89] to extract the underlying human geometry from the neural field. For methods that use the density field, we empirically set the threshold to extract the geometry when applying the Marching Cubes. For our proposed representation “SDF-PDF”, we set the threshold as zero.

Results on the SyntheticHuman datasets: Table III compares our method with [13], [17], [87] in terms of the P2S and CD metrics. [13], [17], [87] and our “NeRF-RVF+”, “NeRF-PDF” representations model the human geometry with the volume density field, while our “SDF-PDF” representation adopt the signed distance field. We empirically set the threshold of volume density to extract the geometry of [13], [17], [87] and our “NeRF-RVF+”, “NeRF-PDF”. Our representation “SDF-PDF” significantly outperforms baseline methods [13], [17], [87] by a margin of at least 0.72 in terms of P2S metric and 0.63 in terms of CD metric. Fig. 5 presents the qualitative comparison.

Results on real data: To further validate the effectiveness of our method, we also perform reconstruction experiments on the real data. Since there are not ground-truth human geometries on real data, we only present the qualitative comparisons. Fig. 6 presents the reconstruction results on the Human3.6 M and MonoCap dataset. The methods are trained on one view on the MonoCap dataset and three views on the Human3.6 M dataset.

D. Ablation Studies

We conduct ablation studies on one subject (S9) of the Human3.6M [12] dataset in terms of the novel pose synthesis performance. First, to analyze the benefit of learning \( \Delta w \), we compare neural blend weight field with SMPL blend weight field. Then, to explore the influence of human pose accuracy, we estimate SMPL parameters from predicted human poses [10], [90] and perform training on these parameters. Finally, we explore the performances of our method under different numbers of video frames and camera views. Ablation studies are performed on the model “NeRF-RVF”, which is the conference version [17]. Tables V, VI, VII, and VIII summarize the results of ablation studies.

Impact of neural residual vector field: Table V shows the quantitative comparisons, which indicate that SMPL blend weight field with neural residual vector field gives better results. To better show the improvement on the SMPL blend weight field, Fig. 7 visualizes the residual vector field \( F_{\Delta w} \) on our reconstructed geometry at the canonical space. The bigger residual has a redder color. We can see that regions of big residual mainly locate on the neck, hand, chest, and pants, which are human-specific details that SMPL cannot describe. The results indicate that our learned \( F_{\Delta w} \) are physically interpretable.

Impact of the human pose accuracy: Table VI compares the models trained with human poses from marker-based and marker-less systems. The results show that more accurate human poses produce better rendering quality. The qualitative comparison is presented in Fig. 8.
Table III

Results of 3D Reconstruction on SyntheticHuman Dataset

|                  | P2S1 | CD1 |
|------------------|------|-----|
|                  | D-NeRF | NB [13] | A-NeRF [87] | NeRF-RVF* | NeRF-PDF | SDF-PDF | D-NeRF | NB [13] | A-NeRF [87] | NeRF-RVF* | NeRF-PDF | SDF-PDF |
| S1               | 3.40  | 1.44 | 1.30  | 4.30  | 1.59  | 0.75   | 2.40   | 1.39  | 1.29  | 2.94  | 1.45  | 0.86   |
| S2               | 3.38  | 1.68 | 1.39  | 4.66  | 1.74  | 0.70   | 2.45   | 1.48  | 1.22  | 3.03  | 1.51  | 0.81   |
| S3               | 3.95  | 1.52 | 1.80  | 4.45  | 1.61  | 0.62   | 2.71   | 1.42  | 1.53  | 3.02  | 1.40  | 0.81   |
| S4               | 4.18  | 1.20 | 1.46  | 2.84  | 1.58  | 0.58   | 2.85   | 1.23  | 1.28  | 2.07  | 1.40  | 0.74   |
| S5               | 1.22  | 1.20 | 3.75  | 2.87  | 1.85  | 0.66   | 1.10   | 1.14  | 3.64  | 2.33  | 1.44  | 0.65   |
| S6               | 1.76  | 1.31 | 1.17  | 2.62  | 1.97  | 0.74   | 1.43   | 1.28  | 1.07  | 2.05  | 1.54  | 0.73   |
| S7               | 1.66  | 1.61 | 1.03  | 2.88  | 2.11  | 0.69   | 1.82   | 1.74  | 1.29  | 2.59  | 2.02  | 0.65   |
| average          | 2.81  | 1.42 | 6.52  | 3.52  | 1.78  | 0.70   | 2.11   | 1.38  | 6.30  | 2.57  | 1.54  | 0.75   |

The first four rows show the results on monocular videos, and the remaining rows present the results on 4-view videos. [87] failed to converge on subject S5.

Fig. 6. 3D reconstruction on the Human3.6 M and MonoCap dataset. The results in the first row are reconstructions from 3-view videos, and the results in the second row are reconstructions from monocular videos.

Table IV

Results of Novel View Synthesis of Training Poses and Novel Poses on MonoCap Dataset in Terms of PSNR and SSIM (Higher is Better)

|                  | Training poses | Novel poses |
|------------------|----------------|-------------|
|                  | PSNR† | SSIM† | PSNR† | SSIM† |
| NB [13]          | 21.76 | 0.872 | 20.83 | 0.854 |
| NHRI [51]        | 21.29 | 0.875 | 20.45 | 0.866 |
| A-NeRF [87]      | 20.52 | 0.845 | 19.53 | 0.828 |
| NeRF-RVF*        | 21.47 | 0.868 | 20.66 | 0.860 |
| NeRF-PDF         | 22.34 | 0.883 | 21.19 | 0.866 |
| SDF-PDF          | 21.89 | 0.885 | 20.88 | 0.869 |

These methods are trained on one view and tested on ten or eleven views. The results show that our proposed methods can generate high-quality novel view and novel pose image synthesis results with even one input view. “NeRF-RVF*” means “NeRF-RVF” [17] with the IDR network [22].

Table V

Comparison of Different Models With and Without Neural Residual Vector Field on the Human3.6 M Dataset

|                  | PSNR | SSIM |
|------------------|------|------|
| w/ Neural residual vector field | 23.72 | 0.886 |
| w/o Neural residual vector field | 21.65 | 0.850 |

Experiments are conducted on the model “NeRF-RVF”, namely the conference version [17].

Table VI

Comparison Between Models Trained With Human Poses From Marker-Based and Marker-Less Pose Estimation Methods on Subject “S9”

|                  | PSNR | SSIM |
|------------------|------|------|
| Marker-based pose estimation | 23.72 | 0.886 |
| Marker-less pose estimation | 22.27 | 0.858 |

Experiments are conducted on the conference version model “NeRF-RVF”.

Table VII

Results of Models Trained With Different Numbers of Video Frames on “S9” of Human3.6 M Dataset

| Frames | PSNR | SSIM |
|--------|------|------|
| 1      | 20.29 | 0.881 |
| 100    | 23.40 | 0.883 |
| 200    | 23.69 | 0.875 |
| 800    | 23.16 | 0.875 |

We perform experiments on the model “NeRF-RVF” [17].

Table VIII

Results of Models Trained With Different Numbers of Camera Views on “S9” of Human3.6 M Dataset

|                  | PSNR | SSIM |
|------------------|------|------|
| 1 view           | 23.81 | 0.877 |
| 2 views          | 24.16 | 0.880 |
| 3 views          | 23.72 | 0.886 |

Experiments are performed on the model “NeRF-RVF” [17].
Fig. 7. Visualization of the residual vector field $F_{\Delta w}$ on the reconstructed geometries of subject “S9”. Red means large residual. Best viewed in color.

Fig. 8. Qualitative results of models trained on poses from marker-less and marker-based systems.

Fig. 9. Comparison of models trained with different numbers of video frames on the subject “S9”.

E. Running Time

For $512 \times 512$ images, the model “NeRF-RVF” takes 1.09 s to render an image on a desktop with an Intel i7 3.7 GHz CPU and a GTX 1080 Ti GPU. Specifically, our implementation takes 0.39 s for predicting the color and density fields, 0.63 s for predicting the blend weight fields, and 0.07 s for volume rendering. Because the number of points sampled along the ray is only 64 and the scene bound of a human is small, the rendering speed of our method is relatively fast.

VI. DISCUSSION

Density field versus Signed distance field: By comparing “NeRF-PDF” and “NeRF-SDF” in Tables I, II and IV, we find that the density field and signed distance field give similar performance in terms of image synthesis on the Human3.6 M, MonoCap, and ZJU-MoCap datasets. In Table III, “SDF-PDF” significantly outperforms “NeRF-PDF” in terms of 3D reconstruction on the SyntheticHuman dataset. The qualitative results in Figs. 5 and 6 also demonstrate that the reconstruction results of “SDF-PDF” are better.

Residual vector field versus Point displacement field: Comparing “NeRF-RVF” with “NeRF-PDF” in Tables I, II and IV, we find that the point displacement field outperforms the residual vector field in terms of image synthesis. In addition, Table III indicates that the point displacement field also achieves better performance on 3D reconstruction.

Network architecture: The model “NeRF-RVF” adopts the network of IDR [22], while the conference version model “NeRF-RVF” uses the network of NeRF [6]. The newly adopted network has a bigger color head than NeRF’s network. Table I shows that the bigger network improves the rendering performance. Detailed network architectures are described in the supplementary material, available online.
Limitations: Combining neural radiance fields with blend weight fields enables us to obtain impressive performances on novel view synthesis and novel pose synthesis. However, our method has a few limitations. (1) The skeleton-driven deformation model [5] cannot express the complex non-rigid deformations of garments. As a result, the performance of our method tends to degrade when reconstructing performers that wear loose clothes. (2) Same to NeRF, our proposed model is trained per-scene, which requires a lot of time to produce animatable human models. Generalizing the networks across different videos and reducing training time is left as future work. (3) Moreover, the rendering time of our model is a bit high. It is could be solved with recent caching-based techniques [91], [92].

VII. CONCLUSION

We introduced a novel dynamic human representation for modeling animatable human characters from multi-view videos. Our method augments a neural radiance field with deformation fields that transform observation-space points to the canonical space. The deformation fields are constructed based on the skeleton-driven deformation framework, where we learn neural blend weight fields to generate observation-to-canonical and canonical-to-observation correspondences. The animatable neural radiance field is learned over the multi-view video with volume rendering and the consistency among blend weight fields. To improve the quality of human modeling, we additionally propose to represent the human geometry as a signed distance field and introduce a neural point displacement field and canonical-to-observation correspondences. The animatable neural radiance field is learned over the multi-view video with volume rendering and the consistency among blend weight fields. To improve the quality of human modeling, we additionally propose to represent the human geometry as a signed distance field and introduce a neural point displacement field to model non-rigid human motions. After training, our method can synthesize free-viewpoint videos of a performer given novel motion sequences. Experiments on the Human3.6 M, MonoCap, ZJU-MoCap, and SyntheticHuman datasets demonstrated that the proposed model achieves state-of-the-art performances on image synthesis and 3D reconstruction.

REFERENCES

[1] J. L. Schonberger and J.-M. Frahm, “Structure-from-motion revisited,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 4104–4113.
[2] K. Guo et al., “The relightables: Volumetric performance capture of humans with realistic relighting,” ACM Trans. Graph., vol. 38, pp. 1–19, 2019.
[3] A. Collet et al., “High-quality streamable free-viewpoint video,” ACM Trans. Graph., vol. 37, pp. 1–13, 2015.
[4] M. Dou et al., “Fusion4D: Real-time performance capture of challenging scenes,” ACM Trans. Graph., vol. 35, pp. 1–13, 2016.
[5] J. P. Lewis, M. Cordner, and N. Feng, “Pose space deformation: A unified approach to shape interpolation and skeleton-driven deformation,” in Proc. Ann. Conf. Comput. Graph. Interactive Techn., 2000, pp. 165–172.
[6] 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 Proc. Eur. Conf. Comput. Vis., 2020, pp. 405–421.
[7] K. Park et al., “NeRFies: Deformable neural radiance fields,” in Proc. Int. Conf. Comput. Vis., 2021, pp. 5845–5854.
[8] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer, “D-NeRF: Neural radiance fields for dynamic scenes,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 10313–10322.
[9] Z. Li, S. Niklaus, N. Snavely, and O. Wang, “Neural scene flow fields for space-time view synthesis of dynamic scenes,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 6494–6504.
[10] H. Joo, T. Simon, and Y. Sheikh, “Total capture: A 3D deformation model for tracking faces, hands, and bodies,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018, pp. 8320–8329.
[11] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, “SMPL: A skinned multi-person linear model,” ACM Trans. Graph., vol. 34, pp. 1–16, 2015.
[12] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, “Human3.6M: Large-scale datasets and predictive methods for 3D human sensing in natural environments,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 7, pp. 1325–1339, Jul. 2014.
[13] S. Peng et al., “Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021, pp. 9030–9059.
[14] M. Habernann, W. Xu, M. Zollhoefer, G. Pons-Moll, and C. Theobalt, “DeepCap: Monocular human performance capture using weak supervision,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 5051–5062.
[15] M. Habernann, L. Liu, W. Xu, M. Zollhoefer, G. Pons-Moll, and C. Theobalt, “Real-time deep dynamic characters,” ACM Trans. Graph., vol. 40, no. 4, 2021, Art. no. 94, doi: 10.1145/3450626.3459749.
[16] S. Peng et al., “Animatable neural implicit surfaces for creating avatars from videos,” 2022, arXiv:2203.08133.
[17] S. Peng et al., “Animatable neural radiance fields for modeling dynamic human bodies,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 14294–14303.
[18] Z. Su et al., “RobustFusion: Human volumetric capture with data-driven view synthesis,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 246–264.
[19] P. Debevec, T. Hawkins, C. Tchou, H.-P. Duiker, W. Sarokin, and M. Sagae, “Acquiring the reflectance field of a human face,” in Proc. Annua. Conf. Comput. Graph. Interactive Techn., 2000, pp. 145–156.
[20] V. Sitzmann, M. Zollhoefer, and G. Wetzstein, “Scene representation networks: Continuous 3D-structure-aware neural scene representations,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2019, pp. 1121–1132.
[21] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger, “Differentiable volumetric rendering: Learning implicit 3D representations without 3D supervision,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 3504–3515.
[22] Y. Li et al., “Multi-view surface reconstruction by disentangling geometry and appearance,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2020, pp. 2492–2502.
[23] R. Natsume et al., “SiCloPe: Silhouette-based clothed people,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 4480–4490.
[24] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li, “PIFp: Pixel-aligned implicit function for high-resolution clothed human digitization,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 2304–2314.
[25] Z. Zheng, T. Yu, Y. Wei, Q. Dai, and Y. Liu, “DeepHuman: 3D human reconstruction from a single image,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 7739–7749.
[26] S. Saito, T. Simon, J. Saragih, and H. Joo, “PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 81–90.
[27] L. Kavan, S. Collins, J. Žára, and C. O’Sullivan, “Skinning with dual quaternions,” in Proc. Symp. Interactive 3D Graph. Games, 2007, pp. 39–46.
[28] G. Pavlakos, L. Zhu, X. Zhou, and K. Daniilidis, “Learning to estimate 3D human pose and shape from a single color image,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018, pp. 459–468.
[29] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik, “End-to-end recovery of human shape and pose,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018, pp. 7122–7131.
[30] N. Kolotouros, G. Pavlakos, and K. Daniilidis, “Convolutional mesh regression for single-image human shape reconstruction,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 4496–4505.
[31] W. Jiang, N. Kolotouros, G. Pavlakos, X. Zhou, and K. Daniilidis, “Coherent reconstruction of multiple humans from a single image,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 5578–5587.
[32] J. Dong, Q. Shuai, Y. Zhang, X. Liu, X. Zhou, and H. Bao, “Motion capture based reconstruction of 3D people models,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018, pp. 1175–1186.
[84] A. Gropp, L. Yariv, N. Haim, M. Atzmon, and Y. Lipman, “Implicit geometric regularization for learning shapes,” in Proc. Int. Conf. Mach. Learn., 2020, pp. 3789–3799.

[85] J. T. Kajiya, “The rendering equation,” in Proc. Annu. Conf. Comput. Graph. Interactive Techn., 1986, pp. 143–150.

[86] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. Int. Conf. Learn. Representations, 2015.

[87] S.-Y. Su, F. Yu, M. Zollhöfer, and H. Rhodin, “A-NeRF: A-NeRF: Articulated neural radiance fields for learning human shape, appearance, and pose,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2021, pp. 12278–12291.

[88] 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 Trans. Graph., vol. 40, no. 6, 2021, Art. no. 219, doi: 10.1145/3478513.3480528.

[89] W. E. Lorensen and H. E. Cline, “Marching cubes: A high resolution 3D surface construction algorithm,” in Proc. Annu. Conf. Comput. Graph. Interactive Techn., 1987, pp. 347–353.

[90] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, “OpenPose: Realtime multi-person 2D pose estimation using part affinity fields,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 1, pp. 172–186, Jan. 2021.

[91] A. Yu, R. Li, M. Tancik, H. Li, R. Ng, and A. Kanazawa, “PlenOctrees for real-time rendering of neural radiance fields,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2021, pp. 5752–5761.

[92] J. Dong received the BE degree in information engineering from Zhejiang University, in 2018 and the PhD degree in the computer science from Zhejiang University, in 2023. He is currently a researcher with Shanghai AI Laboratory. His research interests include 3D human reconstruction and generation.

Xiaowei Zhou received the PhD degree from the Hong Kong University of Science and Technology, after which he was a postdoctoral Researcher with the GRASP Lab, University of Pennsylvania. He is a tenured associate professor of computer science with Zhejiang University, China. His research interests include 3D reconstruction and scene understanding.

Sida Peng received the PhD degree from College of Computer Science and Technology, Zhejiang University, in 2023. He is a research professor of Software with Zhejiang University, China. His research interests include 3D reconstruction, rendering, and 3D generation.

Zhen Xu (Graduate Student Member, IEEE) received the bachelor’s degree in computer science from Zhejiang University, in 2022. He is currently working toward the PhD student in computer science with Zhejiang University, advised by Prof. Xiaowei Zhou. His current research focuses on 3D/4D reconstruction and rendering.

Qianqian Wang received the PhD degree from Cornell Tech, Cornell University. Prior to that she received her B.E. from Zhejiang University. Her research interests include 3D computer vision and graphics. She is a recipient of Google PhD fellowship, in 2022.

Shangzhan Zhang (Graduate Student Member, IEEE) received the bachelor’s degree from Zhejiang University, in 2022. He is currently working toward the master degree in computer science with Zhejiang University, under the guidance of Xiaowei Zhou and Sida Peng. Before pursuing his master degree. His research primarily revolves around 3D/4D reconstruction and generation.

Qing Shuai received the BE degree from Zhejiang University, in 2019. He is currently working toward the PhD student of computer science with Zhejiang University, advised by Dr. Xiaowei Zhou. His research interests include 3D human pose estimation and reconstruction, novel view synthesis, and differentiable rendering problems.

Junting Dong received the BE degree in information engineering from Zhejiang University, in 2018 and the PhD degree in the computer science from Zhejiang University, in 2023. He is currently a researcher with Shanghai AI Laboratory. His research interests include 3D human reconstruction and generation.

Hujun Bao (Member, IEEE) received the graduation from Zhejiang University, in 1987, with the BSc degree in mathematics and the PhD degree in applied mathematics from the same university, in 1993. He is currently a professor in the State Key Laboratory of CAD&CG and College of Computer Science and Technology, Zhejiang University.