Neural Capture of Animatable 3D Human from Monocular Video

Gusi Te[1,2], Xiu Li[2,3], Xiao Li[2], Jinglu Wang[2], Wei Hu[1](✉), and Yan Lu[2](✉)

[1] Peking Univh Asia, Beijing, China
[2] Microsoft Research Asia, Beijing, China
yanlu@microsoft.com, forhuwei@pku.edu.cn
[3] Tencent, Shenzhen, China

Abstract. We present a novel paradigm of building an animatable 3D human representation from a monocular video input, such that it can be rendered in any unseen poses and views. Our method is based on a dynamic Neural Radiance Field (NeRF) rigged by a mesh-based parametric 3D human model serving as a geometry proxy. Previous methods usually rely on multi-view videos or accurate 3D geometry information as additional inputs; besides, most methods suffer from degraded quality when generalized to unseen poses. We identify that the key to generalization is a good input embedding for querying dynamic NeRF: A good input embedding should define an injective mapping in the full volumetric space, guided by surface mesh deformation under pose variation. Based on this observation, we propose to embed the input query with its relationship to local surface regions spanned by a set of geodesic nearest neighbors on mesh vertices. By including both position and relative distance information, our embedding defines a distance-preserved deformation mapping and generalizes well to unseen poses. To reduce the dependency on additional inputs, we first initialize per-frame 3D meshes using off-the-shelf tools and then propose a pipeline to jointly optimize NeRF and refine the initial mesh. Extensive experiments show our method can synthesize plausible human rendering results under unseen poses and views.

1 Introduction

The problem of digital reconstruction, modeling and photo-realistic synthesis of humans from a video sequence such that it can be rendered with any pose from any viewpoint is important, which enables various applications ranging from character animation for games and movies to immersive experience for virtual conferencing. This problem is extremely challenging due to the complicated joint space of human geometry, appearance, and dynamic motion given only RGB videos as observation, especially for monocular videos where multi-view concurrency is unavailable.

G. Te and X. Li—Work done during an internship at Microsoft Research Asia.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-20068-7_16.

S. Avidan et al. (Eds.): ECCV 2022, LNCS 13666, pp. 275–291, 2022.
https://doi.org/10.1007/978-3-031-20068-7_16

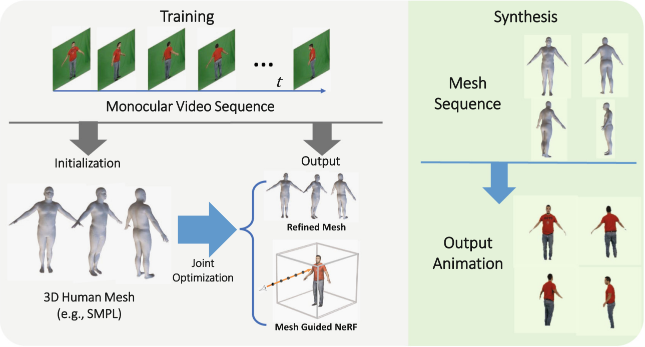


Fig. 1. Left: given a monocular video sequence of human performance with initial posed 3D human with off-the-shelf tools, our method jointly reconstructs a mesh-guided neural radiance field (NeRF) and refined per-frame human mesh. Right: our trained mesh-guided NeRF is rigged with 3D mesh model and enables novel pose and view synthesis.

Because of the difficulty in jointly modeling shape, pose and appearance of 3D humans from monocular videos, many previous approaches focus on solving part of the problem only, such as skeleton-based human pose estimation [4, 10] or parametric 3D model [3,18] based human shape reconstruction [14,32]. These methods exploit sophisticated pose and shape priors and are thus able to partially counteract the geometry ambiguity; however, due to the lack of appearance information, the obtained results might not perfectly align with the input observations in certain frames. Extracted texture based on the estimated surface is usually blurry and cannot be used for photo-realistic synthesis (Fig. 1).

Recently proposed volumetric neural rendering methods, *i.e.* NeRF and its variants [2,19,28], have shown great advances in high-quality free-view synthesis for static objects. NeRF models static objects by an implicit radiance field function with multi-layer perceptron (MLP) networks. Inspired by NeRF, recent works [17,21,24,25] attempt to model 3D humans by conditioning the radiance field on 3D poses/parametric meshes. While promising human reconstruction and view synthesis results have been achieved, these methods only focus on the modeling of conditional radiance field itself and require accurate 3D poses or meshes as a prior. This assumption is often too strong to be fulfilled in practical capture setups, especially with monocular video only.

To this end, we propose a novel paradigm of modeling an animatable 3D human representation from a monocular video sequence of a single person. Our goal is to build a reconstruction pipeline with few non-trivial requirements such as accurate 3D human poses and/or geometry. To achieve this goal, we propose to jointly optimize per-frame human mesh reconstruction and a dynamic neural radiance field (NeRF) which is conditional on mesh information. Given a monocular video sequence as input observations, the optimization process is driven by the re-rendering error on the neural rendering output corresponding to both the NeRF and human poses, which are updated via back-propagation. To constrain

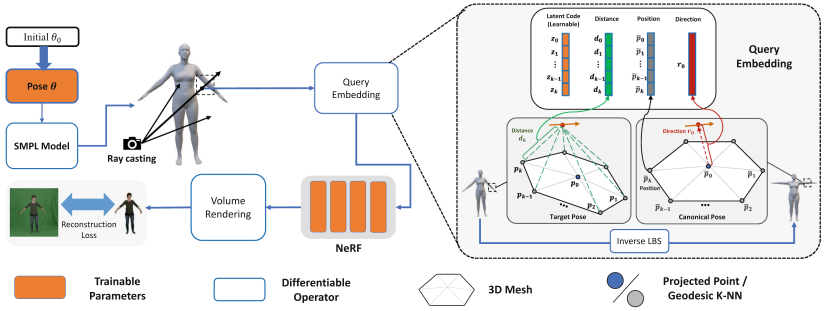


Fig. 2. Our pipeline. A 3D mesh is generated from SMPL model with target pose θ, followed with a mesh-guided NeRF which takes query embedding of 3D points and renders image via volume rendering. The query embedding encodes both surface deformation constraint with information of nearest mesh vertices under rest pose, as well as distance-preserve prior with distance to mesh vertices in a local region under target pose. During training the pose is initialized with off-the-shelf tools and are jointly refined with mesh-guided NeRF.

the optimization space of human mesh, we exploit the widely-used parametric human body model [3], and initialize the optimization with poses provided by monocular pose estimation solutions [14,32] as a starting point. Our joint optimization strategy connects the (previously mangled) 3D geometry estimation problem and NeRF-based appearance optimization problem, and eliminates the requirement of accurate 3D geometry information as a priori, making the modeling pipeline more applicable under monocular video scenarios.

A key property of a good neural representation of humans is that it should have good generalization under unseen human poses after training on limited observations. This is a non-trivial task as previous NeRF-based works for human modeling [24,25] suffer from degraded quality more or less when generalized to unseen human poses. Our observation is that the key for better pose generalization lies in the embedding method of input for querying NeRF. Intrinsically, the dynamic NeRF-based representation of humans can be regarded as a static NeRF under rest pose equipped with 3D volume deformation that is conditioned on the mesh deformation from rest pose to any arbitrary target pose. Thus, a good embedding for querying a dynamic NeRF input under arbitrary poses should "reverse" the pose deformation in an injective way to find the correct point at the static NeRF. As the "correct" deformation mapping is only available on the surface mesh, the reverse deformation at any off-surface region in the space should be constrained with additional priors. Otherwise, the deformation mapping will be distorted and collapsed, thus failing to generalize to unseen poses.

Based on this observation, we propose a new embedding method for querying mesh-guided dynamic NeRF by encoding the input position with its relationship to local nearby surface regions. Specifically, given a query point and a human mesh corresponding to a target pose, we project the query point onto the mesh and find a set of nearest neighbor mesh vertices locally; we then construct the input embedding with distances to these vertices in the **target** space as well as the normalized position of these vertices in the **canonical** space with **rest** human pose, eliminating pose deformation and view transformation.

Out proposed embedding method is able to guide the volume deformation at off-surface points with nearby surface deformation (as we give the inverse-transformed nearest neighbor vertices on mesh). It has two key properties that are essential for improving generalization. First, the embedding is locally based on a nearby small connected region on the guided mesh. The local priors are crucial because they prevent the network from inadvertently relating the output to irrelevant articulated parts, which is known to hurt model generalization to poses unseen during training [21,34]. Second, since we give the distances to all nearest neighboring vertices in the target space, the embedding will encourage a locally distance-preserve prior to restrain the deformation from collapse.

Our method requires only the monocular video of a single person with a fixed camera, which does not rely on dedicated capture devices and/or accurate human pose information. Extensive experimental results demonstrate the superiority of our model on a variety of data that exhibit various human shapes and poses. To summarize, our contributions are as follows:

- We propose a novel paradigm for building a neural human representation that can be rendered in unseen poses and views with monocular video inputs.
- We propose a novel input embedding representation for querying mesh-guided NeRF which improves the generalization ability on novel poses.
- We develop a pipeline for joint optimization of 3D human meshes and mesh-guided dynamic NeRF supervised by the reconstruction loss only.

2 Related Works

Human Reconstruction. The problem of digital reconstruction of humans is a long-standing problem in computer vision and computer graphics. Traditional methods usually achieve high quality with complicated capture setups such as multi-view capture studio [8,35,36] or RGB-D camera arrays [30,33]. To reduce capture efforts, recent methods leverage deep neural networks to directly reconstruct 3d humans from even single images [7,14,20,27]. These methods often estimate output coefficients of parametric models of 3D human shape and poses [18]. The parametric model of 3D humans is often constructed from a large database of scanned shapes of different humans in a variety of poses and the rigged with a pre-defined skeleton to animate the human mesh.

Neural 3D Representations. Recently, neural representation of 3D scenes has attracted considerable attention in the literature [2,5,6,19,22,28]. These methods exploit a neural network (usually multi-layer perceptrons) to represent implicit fields such as signed distance functions for surface or volumetric radiance fields, thus inherently encoding 3D information in a view-consistent manner. Among those neural representations, NeRF [19] (and its variants) has surpassed previous state-of-the-art methods on novel view synthesis tasks for static objects. Some works also extended NeRF to handle general space-time dynamic scenes [23,26,31]. Our method targets extending NeRF to model dynamic representation of 3D humans with the help of parametric 3D body mesh models.

Rigging NeRF. A prevalent approach for representing dynamic humans with NeRF is to rig NeRF with articulated models. Common articulation choices are 3D pose skeletons [21,29] and parametric 3D mesh models [9,17,24,25]. Our method utilizes a parametric 3D mesh model [3] for articulation. While we are similar to previous and concurrent works [17,21,24,29] by sharing the same goal of modeling dynamic human body with articulated NeRF representation, our method differs them in two aspects. First, we attempts to simplify the input to monocular video input as opposed to multi-view video inputs [17,24] and relax the dependence on accurate 3D geometry input [21] a priori. Second, we propose a new embedding method for querying articulated dynamic NeRF with locality and distance-preserving constraints. Noguchi et al. [21] proposed to learn a most relevant articulated part for any given query point. The concurrent work of Su et al. [29] propose a similar framework with joint-optimization of NeRF and human pose, using the skeleton as the human shape representation and directly relates the input query to all articulated skeleton joints. We focus on improving the generalization ability for NeRF-based animatible 3D human reconstruction with novel embedding designs. Our method preserves locality via nearest-neighbor projection, and encourages locality distance-preserving to avoid collapse of deformation in the whole volume.

3 Method

Given a monocular video sequence $\{\mathbf{I}_i\}_{i=1}^{K}$ as input, we aim to construct a neural human representation that encodes both appearance and geometry knowledge and can be rendered under an arbitrary pose θ. In particular, we model our representation with a neural radiance field (NeRF). Our NeRF is dynamically controlled by an underlying parametric mesh model (Sect. 3.1). Given an observation-space pose, the mesh surface is deformed from its rest pose correspondingly. We design a novel query embedding (Sect. 3.2) for the input which encodes both information of surface deformation and addition constraints. Based on the proposed mesh-guided NeRF, we propose an analysis-by-synthesis method to jointly estimate pre-frame 3D mesh from the input video and train NeRF (Sect. 3.3), using off-the-shelf tools for mesh initialization.

3.1 Mesh-Guided NeRF

In NeRF, the rendered color $\bar{\mathbf{C}}(u, v)$ at image pixel (u, v) is generated by querying and blending the radiance along the corresponding camera ray according to the volume density value:

$$\bar{\mathbf{C}}(u, v) = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma_i \delta_i))\mathbf{c}_i, \tag{1}$$

where

$$T_i = \exp(-\sum_{j=1}^{i-1}(-\sigma_j \delta_j))), \tag{2}$$

and

$$(\mathbf{c}_i, \sigma_i) = F(\mathbf{x}_i). \tag{3}$$

$\mathbf{c}_i \in \mathcal{R}^3$ and σ_i are the color and volume density of the i-th sampled point $\mathbf{x}_i$ along the ray direction. $F(\mathbf{x})$ is usually parameterized with an MLP network.

We extend NeRF to handle the dynamic, articulated human body with a mesh-based parametric 3D model SMPL [18]. An SMPL model $S(\theta, \beta)$ takes a human 3D pose θ of skeleton joint rotations as well as a low-dimensional feature vector of human shape as input and outputs a 3D mesh. As we mainly focus on synthesizing humans under different poses, we omit the shape β afterwards.

Formally, given a pose input θ, the radiance color $\mathbf{c}(\mathbf{x})$ and volume density of our mesh-guided NeRF at point x is computed as follows:

$$(\mathbf{c}(\mathbf{x}), \sigma(\mathbf{x})) = F_\Phi(q(\mathbf{x}, S(\theta))), \tag{4}$$

where the query embedding q is the most important part as it directly relates the output of NeRF with the underlay deformable mesh, as we will discuss next.

3.2 Query Embedding for NeRF

The input of NeRF for querying radiance value at point $\mathbf{x}$ is given by its 3D location (x, y, z) and 2D viewing direction θ, ϕ in the world space. A natural extension of input querying for the dynamic scene is to define a deformation field that transforms observation-space points to rest space. Directly estimating a general deformation field together with the NeRF, as in [23,26,31], is highly ill-posed and prone to local minima. Inspired by [17,24], we leverage the deformable SMPL model as the human prior to guide our transformation for input queries. The underlay SMPL model defines reasonable deformation fields on its surface; however, a radiance field from NeRF is defined on full 3D volume, and we still need to determine the deformation on unconstrained off-surface points. Naively projecting off-surface points to its nearest vertex point on the mesh is not optimal because the off-surface deformation will be collapsed, as illustrated in Fig. 3.

We address this issue from another perspective: instead of inputting an inverse-transformed point with an explicitly defined deformation field for querying NeRF, we construct a query embedding of the input point which encodes two types of information: (1) information that guides how the deformation field should roughly be (denoted as *Deformation Guidance*), and (2) priors that prevent the deformation field from collapsed local minima (denoted as *Deformation Priors*). The NeRF then implicitly learns a radiance field based on the input embedding. Figure 2 illustrates our design of query embedding.

Deformation Guidance. Our deformation guidance is based on the underlay SMPL model. For the SMPL model, the transformation relationship between a

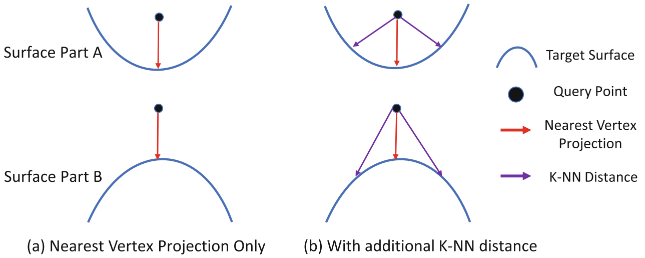

Surface Part A
Surface Part B
Target Surface
Query Point
Nearest Vertex Projection
K-NN Distance
(a) Nearest Vertex Projection Only
(b) With additional K-NN distance

Fig. 3. An illustration of our distance-preserved query embedding. (a) Naively embedding the query with nearest neighbor vertex on mesh (the red line), leads to indistinguishable embedding of different surface deformation patterns. (b) With additional geometric K-NN distance information (purple lines), different deformation patterns are clearly separated.

canonical-space surface point $\mathbf{v}$ and its observation space counterpart $\mathbf{v}'$ is given by the linear blend skinning (LBS) algorithm [15]:

$$\mathbf{v}' = (\sum_{k=1}^{K} w(\mathbf{v})_k G_k)\mathbf{v}, \tag{5}$$

where K is the number of human parts, $G_k \in SE(3)$ is the transformation matrix of the k-th part on the human skeleton, and $w(\mathbf{v})$ is the blend weight.

Intuitively, the guidance information from the SMPL model should neither be too global such that the network inadvertently relates the output to irrelevant articulated parts [21,34], nor collapse to a single nearest neighboring point as the deformation field will remain unconstrained (Fig. 3). To this end, we build the deformation guidance part of the input query with the nearest projected vertex on the mesh as well as the k-nearest adjacent vertices of the projected vertex in rest space via inverse LBS as:

$$q_g(\mathbf{x}) = (\mathbf{x}_{dir}, \mathbf{v}_0, \mathbf{v}_1, ..., \mathbf{v}_k), \tag{6}$$

where $\mathbf{v}_k = (\sum_{l=1}^{L} w(\mathbf{v}_k)_l G_l)^{-1}\mathbf{v}'_k$ and $\mathbf{v}'_k$ is the k-th nearest neighboring mesh point in the observation space. Note that, we additionally give the relative direction from query point $\mathbf{x}$ to its projected point $\mathbf{v}_0$:

$$\mathbf{x}_{dir} = \mathcal{R}((\sum_{k=1}^{K} w(\mathbf{v})_k G_k)^{-1}) \frac{\mathbf{v}_0 - \mathbf{x}}{\|\mathbf{v}_0 - \mathbf{x}\|_2}. \tag{7}$$

Here $\mathcal{R}$ denotes the rotational part of the transformation matrix.

Deformation Priors. Our deformation guidance embedding q_g itself is based on mesh surface only and insufficient to ensure a well-defined deformation field in the whole volume. We therefore provide an additional part to the query embedding by equipping the input query with the Euclidean distances to its nearest points in the observation space:

$$q_p(\mathbf{x}) = (d_0, d_1, ..., d_k), \tag{8}$$

where $d_k = \|\mathbf{v}'_k - \mathbf{x}\|_2$. Using the distance in the observation space is important as such information preserves the local difference under different poses and leads to a distance-preserved deformation field.

Appearance Latent Code. To better capture the geometry and appearance detail which cannot be captured by surface mesh deformation, we additionally provide a learnable latent code l_k defined on each mesh vertex:

$$q_a(\mathbf{x}) = (\mathbf{l}_0, \mathbf{l}_1, ..., \mathbf{l}_k). \tag{9}$$

The complete query embedding for the NeRF input is generated by feeding the concatenation vectors into a tiny 3-layer MLP network ψ:

$$q(\mathbf{x}) = \psi(\gamma(q_g(\mathbf{x})), \gamma(q_p(\mathbf{x})), q_a(\mathbf{x})), \tag{10}$$

where γ denotes positional encoding as used in the original NeRF [19].

3.3 Joint Mesh Estimation and NeRF Training

Training the mesh-guided NeRF from monocular video input requires paired data of input frames $\{\mathbf{I}_i\}$ and human mesh $\{\mathbf{M}_i\}$. State-of-the-art monocular video based human mesh reconstruction methods such as [13,14] produce plausible results for human mesh estimation; however, they are still not accurate enough for training our NeRF as non-aligned mesh part to the image will give incorrect guidance and make the NeRF over-fitting to misaligned training poses. Hence we opt to use the plausible mesh estimates provided by prior solutions as initialization, and jointly fine-tune the mesh with NeRF training. Practically, we choose to optimize the pose parameter θ^i for each training frame instead of per-vertex mesh offset, as it gives us enough capability to refine mesh-image misalignment without too much flexibility that overfits to local minima.

Training Objective. Our training is guided by the reconstruction error between the mesh-guided NeRF and the ground-truth frames over the whole video sequence as well as a regularization term penalizing too large deviation from the initial pose estimation θ_0:

$$L = \sum_i \sum_{u,v} L_i(u, v) + \lambda_p \sum_i \|\theta^i - \theta_0^i\|_2^2, \tag{11}$$

and

$$L_i(u, v) = \|\bar{\mathbf{C}}(u, v) - \mathbf{I}_i(u, v)\|_2^2, \tag{12}$$

where $\mathbf{I}_i(u, v)$ is the ground truth pixel value at (u, v) from the i-th frame. $\bar{\mathbf{C}}(u, v)$ is computed using Eq. 1, Eq. 4 and the proposed query embedding (Eq. 10).

4 Experiments

4.1 Experimental Setup

Datasets. We conduct experiments on different datasets as follows:

- People-Snapshot [1]: This dataset contains 24 subjects with monocular videos performing turning around. Among them, we choose female-1-casual, female-3-casual, male-1-sport and male-9-plaza for training. We remove the background of the video frames with ground truth silhouettes provided and resize the video to half-size (1080p → 540p). An initial mesh is provided in the data.
- DoubleFusion [33]: This dataset contains only a sequence of one man, where the actor performs more complex actions while turning around. Thus, we consider it not suitable for a quantitative benchmark and only use it to show qualitative comparisons on novel pose synthesis. The initial mesh is provided in the dataset using additional depth information.
- ZJU-MoCap [25]: This dataset contains multi-view video sequences of 9 objects with 21 cameras. We choose a single view (subject 313 and subject 386 from camera 7) for training.
- Human3.6M [11]: This dataset consists of a large number of 3D human poses and corresponding multi-view video sequences. We follow the same protocol as [29], extracting every 64th frame of the videos. We train the model on the subject 9 and subject 11. For each video, we select camera 2 as the input view and employ SPIN [14] to estimate the initial mesh from video frames.

Network Structure. The network ψ in the query embedding module is implemented with a 3-layer MLP with 128 channels. The NeRF network ϕ is composed of an 8-layer MLP with 256 channels. In the position embedding module, we implement the tiny 3-layer MLP ψ with 128 channels, and the NeRF module ϕ for rendering is composed of 8-layer MLP with 256 channels. We apply a positional encoding of 10 frequencies to query embedding features except latent codes.

Training Details. We utilize Adam optimizer [12] with learning rate of $1e{-}4$ for optimizing NeRF and latent code. The learning rate of body poses is set to $5e{-}4$ and λ_p is set to 2.0. For volumetric rendering we employ the coarse-to-fine ray sampling strategy of [19]. We also constrain the sampled rays to be more focused on the human part in the image by sampling rays within the 1.2× padding bounding box of 2D keypoints with 70% probability, and randomly sampled in the whole image with 30% probability. Each sampled ray is discreted within $[z_{near} - 0.04, z_{far} + 0.04]$, where z_{near} and z_{far} denote the nearest and farthest ray-point intersection with body mesh, respectively. Our model is trained with a single Nvidia Tesla V100 32 GB GPU, and the training approximately takes 60 h to converge. For datasets without background mask available, we either apply an off-the-shelf matting algorithm [16] or jointly model the background during training. Please check the supplemental materials for details.

Evaluation Metrics. Peak-Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) are used to evaluate image quality.

Table 1. Ablation studies on (a) type of direction, (b) type of distances and (c) type of neighborhood selection for embedding construction.

	SSIM ↑		PSNR ↑	
	Training	Novel	Training	Novel
w/o distance	0.935	0.906	29.18	27.51
canonical distance	0.926	0.916	29.06	27.48
observation distance	**0.980**	**0.973**	**35.87**	**34.75**

(a) Impact of distance embedding.

	SSIM ↑		PSNR ↑	
	Training	Novel	Training	Novel
w/o direction	0.942	0.923	30.52	28.15
w/o inverse	0.959	0.921	32.14	28.12
full	**0.980**	**0.972**	**35.87**	**34.75**

(b) Impact of direction embedding.

	SSIM ↑		PSNR ↑	
	Training	Novel	Training	Novel
Euclidean, 2 neighbors	0.950	0.935	32.35	30.69
Geodesic, 2-hop neighbors	0.962	0.954	32.74	31.97
Geodesic, only nearest neighbor	0.961	0.938	31.87	30.07
Geodesic 1-hop neighbors	**0.980**	**0.972**	**35.87**	**34.75**

(c) Impact of neighborhood selection.

4.2 Ablation Studies

To validate the influence of our proposed query embedding, we conduct the ablation study on the People-Snapshot dataset and report quantitative results on both training and test (unseen) poses, from the following aspects:

Neighborhood Range: As we have discussed in Sect. 3.2, the deformation guidance from the SMPL model should be neither too global nor too local. We verified this by conducting training with different ranges of mesh neighborhood. The results are shown in Table 1c. Either increasing range (*2-hop neighbors*) or *only nearest neighbor* projected point leads to degraded performance, both for training and novel poses. We also test a variant of our method by sampling K-NN point based on Euclidean distance (*spatial K-NN*) instead of geodesic distance. The results are also degraded as it fails to aware human part connectivity (*i.e.* two adjacent points in Euclidean space might belong to distinct human parts).

Distance Prior: We validate the importance of distance information in Table 1a. We remove the distance feature in the *w/o distance* model, and substitute rest-pose distance for observation-pose distance in the *canonical distance* model. Obviously, without distance information, the results are significantly degraded and the difference between training and novel poses is increased.

Relative Direction: The impact of relative direction embedding is demonstrated in Table 1b, where *w/o direction* denotes embedding without direction, and *w/o inverse* denotes embedding direction in observation space. It is worth noting that the *w/o inverse* greatly reduces the generalization on novel poses.

Pose Refinement: Our joint pose refinement with NeRF training is crucial when the initial mesh is not accurate enough. To validate this, we conduct experiments on both Human3.6M and People-Snapshot dataset. The People-Snapshot dataset has provided an initial mesh that is rather reasonable; yet, we

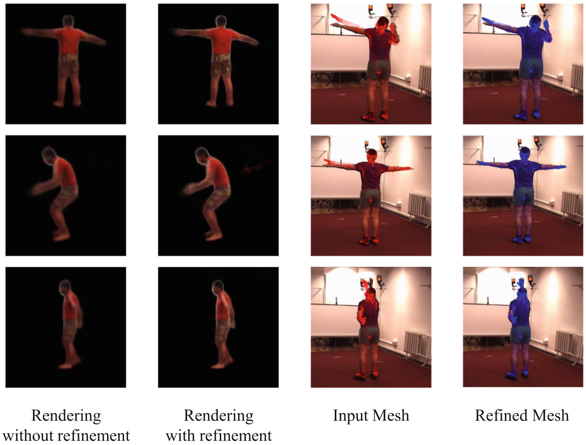


Fig. 4. Qualitative comparison between original and optimized mesh. The final result corrects the initial human mesh, e.g., the alignment error on the arms.

Table 2. The effect of using joint pose refinement.

Dataset	Method	SSIM ↑	PSNR ↑
Human3.6M	w/ refinement	**0.978**	**31.51**
	w/o refinement	0.951	29.04
People-Snapshot	w/ refinement	**0.972**	**34.75**
	w/o refinement	0.969	32.99

Table 3. Quantitative comparison with AniNeRF and A-NeRF.

	ZJU-Mocap		Human3.6M			People-snapshot	
	AniNeRF	Ours	AniNeRF	A-NeRF	Ours	AniNeRF	Ours
SSIM ↑	0.758	**0.768**	0.865	**0.928**	0.912	0.948	**0.973**
PSNR ↑	23.75	**25.01**	23.44	**27.45**	27.11	29.11	**34.75**

still observe minor artifacts without pose refinement and our joint training further improves the result, both quantitatively (Table 2) and qualitatively (Fig 4 and Fig. 5).

4.3 Comparsions

As there is very few (formally peer-reviewed and published) NeRF-based work that shares the same succinct **monocular** inputs with **mesh-based** geometry proxy as ours, we compare with the following methods:

AniNeRF (ICCV 2021). AniNeRF [24] is NeRF-based method for dynamic human modeling. AniNeRF also uses mesh as geometry guidance but requires more strict input requirements of multi-view video input. It produces high quality results with typically 3 to 4 synchronized views. For a fair comparison, we follow the same single view setting and training data to re-train AniNeRF, and report the comparison results in Fig. 7. We emphasize that this experiment setup with monocular input is **not** for producing best-quality results, but to demonstrate the challenge of monocular video scenario as well as the benefit of our proposed method. Compared with AniNeRF, our method generates complete skin and cloth, whereas AniNeRF is unable to model the whole body with limited view. The quantitative result reported in Table 3 also shows our method outperforms AniNeRF under the same settings. We also refer to the supplemental material for a comparsion to NeuralBody [25], the precursor method of AniNeRF.

A-NeRF (NeurIPS 2021). A-NeRF [29] is a recent work for modeling 3D human with NeRF using monocular video input. A-NeRF exploits joint optimization of NeRF with human skeletons. An apple-to-apple comparsion with A-NeRF is hard as it differs from our method in many implementation aspects which affects the result quality, e.g., from the underly parametric body representation (skeleton-based v.s. mesh-based) to the backbone capacities. Nevertheless, our result on the Human3.6M dataset is quantitatively comparable with A-NeRF (Table 3).

Non-NeRF methods. Regarding non-NeRF methods, we also compare our method with a SMPL-model based method, VideoAvatar [1]. The qualitative results are shown in Fig. 6. Given the same monocular video as input, the NeRF-based method generates results with more natural and realistic color effects.

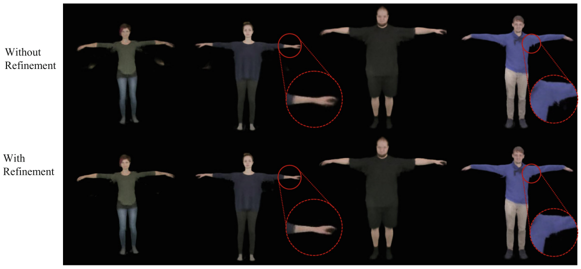


Fig. 5. The effect of pose refinement on People-Snapshot dataset. Jointly refinement contributes to clearer geometry and eliminates outliers. The improvement brought by refinement is enlarged in red.

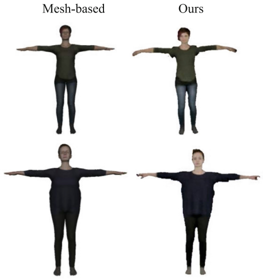
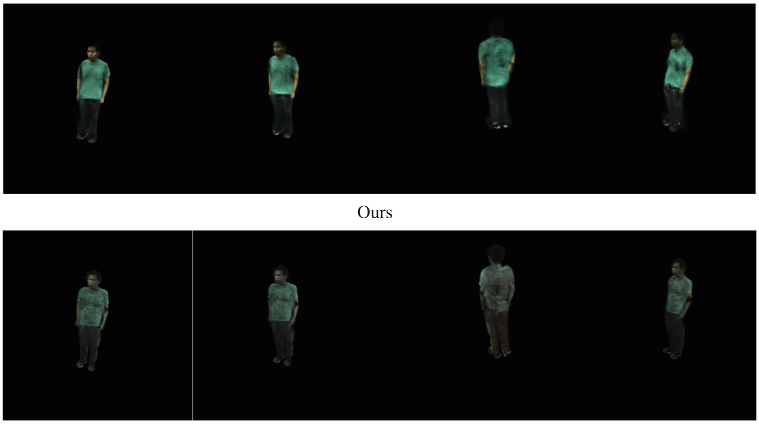


Fig. 6. A qualitative comparison with mesh-based method VideoAvatar.

4.4 Applications

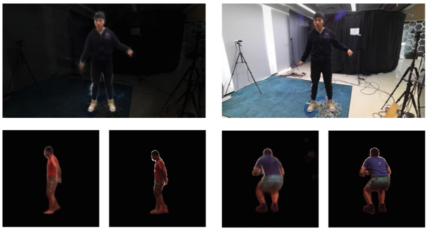
Novel Pose Synthesis. Our trained representation enables character animation from novel unseen poses. We evaluate our generalization ability by comparing testing data and our rendering driven by the same set of unseen poses on the People-Snapshot and DoubleFusion dataset. The qualitative result is depicted in Fig. 8a. Our model successfully disentangles background and foreground pixels and veritably reconstructs the human body in the Doublefusion dataset (First row). As for side and back view, our model still generates images of high quality



Ours

Animatable NeRF

Fig. 7. Qualitative comparison of Ani-NeRF [24] and ours under novel view. Both methods are trained with single view sequence.



(a) Qualitative results

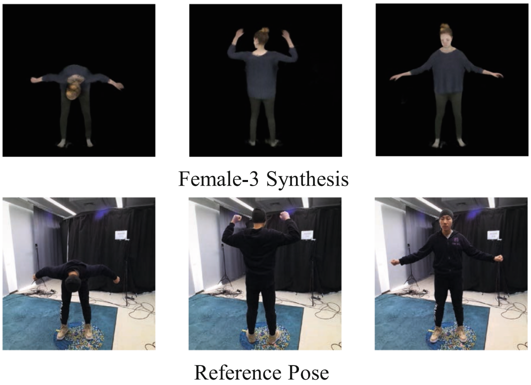
Metric	Human3.6M[11]		People-Snapshot[1]
	S9	S11	Female-1
SSIM ↑	0.978	0.972	0.973
PSNR ↑	31.51	29.70	34.75

(b) Quantitative results

Fig. 8. Qualitative and quantitative results of **novel** pose synthesis on multiple datasets. (a) Top row: novel pose rendering (left) and ground truth (right) on DoubleFusion. Bottom row: rendering (odd column) and ground truth (even column) on Human3.6M. (b) Quantitative results of novel pose synthesis on Human3.6M and People-Snapshot dataset.

as shown in the Human3.6M dataset (Second row). We also provide quantitative results in Fig. 8b on the People-Snapshot and Human3.6m datasets.

Pose Retargeting. The generalization ability of our model is further evaluated by pose retargeting experiments. The results are shown in Fig. 9, where the driven poses derive from the Doublefusion dataset and training body comes from the People-snapshot dataset. We observe that our model generates realistic human bodys with various poses, which demonstrates the generalization of the proposed methods. We refer to the supplemental material for more novel pose synthesis results, including animation videos.



Female-3 Synthesis

Reference Pose

Fig. 9. Human animation driven by Doublefusion poses. The synthetic human is trained on the People-Snapshot dataset.

5 Conclusion

We presented a new method for building animatable neural 3D human representations from only monocular video inputs. Our representation is based on dynamic Neural Radiance Field guided by parametric 3D human meshes. We designed a novel input query embedding of the mesh-guided NeRF. We train the representation by we first initialize per-frame 3D meshes using off-the-shelf tools and then joint optimizing the 3D mesh and dynamic NeRF. The learned neural representation can generalize well to unseen views and poses.

Limitations. Our method is not without limitations. The input embedding of our querying is related to a local region on the mesh surface with a restricted reception field; thus the joint optimization might fail if the initial pose has deviated too much from the ground truth. Due to resolution constraint and the expressiveness of the mesh model we used, our method is still straggling at recovering high-resolution details such as human faces.

Future Work. For future works, we plan to explore different kinds of deformation priors and their effects on rigging dynamic NeRF, improving our performance with sharp details, and extending to general, non-articulated dynamic objects.

Acknowledgements. We would like to thank all the reviewers for their constructive feedback.

References

1. Alldieck, T., Magnor, M., Xu, W., Theobalt, C., Pons-Moll, G.: Video based reconstruction of 3D people models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8387–8397, June 2018. https://doi.org/10.1109/CVPR.2018.00875. CVPR Spotlight Paper

2. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-NeRF: a multiscale representation for anti-aliasing neural radiance fields. arXiv preprint arXiv:2103.13415 (2021)
3. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it SMPL: automatic estimation of 3D human pose and shape from a single image. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 561–578. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_34
4. Cao, Z., Hidalgo, G., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: realtime multi-person 2D pose estimation using part affinity fields. IEEE Trans. Pattern Anal. Mach. Intell. **43**(1), 172–186 (2019)
5. Chen, X., Zheng, Y., Black, M.J., Hilliges, O., Geiger, A.: SNARF: differentiable forward skinning for animating non-rigid neural implicit shapes. In: International Conference on Computer Vision (ICCV) (2021)
6. Deng, Y., Yang, J., Tong, X.: Deformed implicit field: modeling 3D shapes with learned dense correspondence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10286–10296 (2021)
7. Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., Tong, X.: Accurate 3D face reconstruction with weakly-supervised learning: from single image to image set. In: Proceedings of IEEE Computer Vision and Pattern Recognition Workshop on Analysis and Modeling of Faces and Gestures (2019)
8. Dou, M., et al.: Fusion4D: real-time performance capture of challenging scenes. ACM Trans. Graph. (ToG) **35**(4), 1–13 (2016)
9. Guo, Y., Chen, K., Liang, S., Liu, Y., Bao, H., Zhang, J.: AD-NeRF: audio driven neural radiance fields for talking head synthesis. arXiv preprint arXiv:2103.11078 (2021)
10. He, Y., Yan, R., Fragkiadaki, K., Yu, S.I.: Epipolar transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7779–7788 (2020)
11. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3. 6m: large scale datasets and predictive methods for 3D human sensing in natural environments. IEEE Trans. pattern Anal. Mach. Intell. **36**(7), 1325–1339 (2013)
12. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
13. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: video inference for human body pose and shape estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5253–5263 (2020)
14. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2252–2261 (2019)
15. Lewis, J.P., Cordner, M., Fong, N.: Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, pp. 165–172 (2000)
16. Lin, S., Yang, L., Saleemi, I., Sengupta, S.: Robust high-resolution video matting with temporal guidance. arXiv preprint arXiv:2108.11515 (2021)
17. Liu, L., Habermann, M., Rudnev, V., Sarkar, K., Gu, J., Theobalt, C.: Neural actor: Neural free-view synthesis of human actors with pose control. arXiv preprint arXiv:2106.02019 (2021)
18. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: a skinned multi-person linear model. ACM Trans. Graph. (Proc. SIGGRAPH Asia) **34**(6), 248:1–248:16 (2015)

19. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: representing scenes as neural radiance fields for view synthesis. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 405–421. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_24
20. Natsume, R., et al.: SiCloPe: silhouette-based clothed people. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4480–4490 (2019)
21. Noguchi, A., Sun, X., Lin, S., Harada, T.: Neural articulated radiance field. arXiv preprint arXiv:2104.03110 (2021)
22. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: DeepSDF: learning continuous signed distance functions for shape representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 165–174 (2019)
23. Park, K., et al.: Nerfies: deformable neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5865–5874 (2021)
24. Peng, S., et al.: Animatable neural radiance fields for human body modeling. arXiv preprint arXiv:2105.02872 (2021)
25. Peng, S., et al.: Neural body: implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In: CVPR (2021)
26. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-NeRF: neural radiance fields for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10318–10327 (2021)
27. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2304–2314 (2019)
28. Sitzmann, V., Martel, J., Bergman, A., Lindell, D., Wetzstein, G.: Implicit neural representations with periodic activation functions. In: Advances in Neural Information Processing Systems 33 (2020)
29. Su, S.Y., Yu, F., Zollhoefer, M., Rhodin, H.: A-NeRF: surface-free human 3D pose refinement via neural rendering. arXiv preprint arXiv:2102.06199 (2021)
30. Su, Z., Xu, L., Zheng, Z., Yu, T., Liu, Y., Fang, L.: RobustFusion: human volumetric capture with data-driven visual cues using a RGBD camera. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020, Part IV. LNCS, vol. 12349, pp. 246–264. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58548-8_15
31. Xian, W., Huang, J.B., Kopf, J., Kim, C.: Space-time neural irradiance fields for free-viewpoint video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9421–9431 (2021)
32. Xiang, D., Joo, H., Sheikh, Y.: Monocular total capture: posing face, body, and hands in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10965–10974 (2019)
33. Yu, T., et al.: DoubleFusion: real-time capture of human performances with inner body shapes from a single depth sensor. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7287–7296 (2018)
34. Zeng, A., Sun, X., Huang, F., Liu, M., Xu, Q., Lin, S.: SRNet: improving generalization in 3D human pose estimation with a split-and-recombine approach. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12359, pp. 507–523. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58568-6_30

35. Zhang, Y., Li, Z., An, L., Li, M., Yu, T., Liu, Y.: Lightweight multi-person total motion capture using sparse multi-view cameras. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5560–5569 (2021)
36. Zheng, Y., et al.: DeepMultiCap: performance capture of multiple characters using sparse multiview cameras. arXiv preprint arXiv:2105.00261 (2021)