Structured Local Radiance Fields for Human Avatar Modeling

Zerong Zheng[1], Han Huang[2], Tao Yu[1], Hongwen Zhang[1], Yandong Guo[2], Yebin Liu[1]

[1]Department of Automation, Tsinghua University [2]OPPO Research Institute

Abstract

It is extremely challenging to create an animatable clothed human avatar from RGB videos, especially for loose clothes due to the difficulties in motion modeling. To address this problem, we introduce a novel representation on the basis of recent neural scene rendering techniques. The core of our representation is a set of structured local radiance fields, which are anchored to the pre-defined nodes sampled on a statistical human body template. These local radiance fields not only leverage the flexibility of implicit representation in shape and appearance modeling, but also factorize cloth deformations into skeleton motions, node residual translations and the dynamic detail variations inside each individual radiance field. To learn our representation from RGB data and facilitate pose generalization, we propose to learn the node translations and the detail variations in a conditional generative latent space. Overall, our method enables automatic construction of animatable human avatars for various types of clothes without the need for scanning subject-specific templates, and can generate realistic images with dynamic details for novel poses. Experiment show that our method outperforms state-of-the-art methods both qualitatively and quantitatively.

Figure 1. **Example results produced by our method.** Our method can learn animatable human avatars with various cloth topologies and realistic dynamic details. Top row: driving video, from which the animation poses are extracted. Bottom two rows: animation results rendered from the front and the back view.

1. Introduction

Animatable human avatar modeling is of great importance in many applications such as content creation and entertainment, and virtual characters have become ubiquitous in our lives with the rise of computer graphics in movies and games. Traditional methods for high-quality human avatar reconstruction are often costly and tedious, due to the difficulties in modeling the complex dynamics of clothes. Besides, they typically presume the availability of a subject-specific template [22] and its accurate registration to the input frames [6, 78], which are difficult to acquire in practice.

With the rapid development in computer vision in the past ten years, researchers have started to explore the possibility of automatic human avatar reconstruction without pre-scanning efforts. Pioneer studies deformed a statistical human body template (e.g., SMPL [40]) to model the clothed human geometry and appearance [2–4]. Neural texture maps and image-to-image networks are later adopted to achieve photo-realistic rendering [36, 37, 58, 65]. Recently, neural radiance representations, which implicitly encode shape and appearance using neural networks, are also applied in pursuit of higher-fidelity results [35, 49, 54]. These methods typically define the radiance field in a canonical pose, and warp it to live poses using linear blending skinning (LBS) under the guidance of the SMPL surface.

Despite the differences in the representations inside the aforementioned approaches, we find that there is one thing in common: they all heavily rely on the skeleton or the surface of SMPL model for cloth motion modeling. This is apparent in methods based on the SMPL topology, either using traditional texture maps [2–4] or neural textures [36, 37, 58, 65]. Even in state-of-the-art methods based on implicit fields [35, 49, 54], researchers still assumed that skin motions can be propagated to approximate the cloth deformations, which, unfortunately, only holds for tight-fitting clothes. When applying these methods to loose clothes, articulation motions based on solely body joints cannot express the complete information about the wrinkles and non-rigid deformations. Some methods learned to directly regress cloth deformations from body pose configurations [35]; however, the complexity gap between body poses and cloth details results in a one-to-many mapping problem, leading to under-fitting issues where the network learns averaged, blurry appearance. Suffering from this fatal limitation, no methods have demonstrated animatable human characters wearing skirts or dresses so far.

To overcome this limitation and fill the void, we propose a new representation for clothed human characters. Our representation is built upon neural radiance fields [46], or NeRF in short, for its excellent performance in learning the appearance of static scenes. To extend NeRF for dynamic character modeling, we break a global NeRF into a set of *structured local radiance fields*, which are attached to the pre-defined nodes on the SMPL model. Each local radiance field is responsible for representing the shape and appearance in the local space around its corresponding node. The local radiance fields can be driven by the body skeleton, while having their own residual movements to represent the non-rigid deformation of garments. Furthermore, each radiance field is conditioned on a dynamic detail embedding, which encodes the high-frequency dynamic details that cannot be modeled via node translation. In this way, our representation decomposes the cloth deformations in a coarse-to-fine manner: the coarsest level is the skeleton motion, the middle level is the residual movements of the local radiance fields, and the finest level is the time-varying details inside each radiance field.

However, employing such a representation for avatar modeling is not straight-forward as the node-related variables (*i.e.*, the node residual translations and the dynamic detail embeddings) are difficult to acquire in practice. Although we can obtain these variables for training frames through naive optimization with image evidence, it remains unclear how to compute them for unseen poses. Alternatively, one can train a network that directly regresses these variables from body poses, but this will result into the aforementioned under-fitting issues due to information deficiency [6]. In order to achieve a balance between data fitting and generalization, we draw inspiration from [6] and learn the node-related variables in a conditional generative latent space. Specifically, we introduce a tiny conditional variational auto-encoder (cVAE) [68] for each local radiance field. Conditioned on the pose parameters, the cVAE decoders convert the latent bottlenecks into node-related variables. For the input of the cVAE encoder, we find that the time stamp [16, 57, 77] is an effective option, because it is simple, distinguishable, and naturally guarantees the temporal smoothness of the node-related variables thanks to the low-frequency bias in MLPs [71]. Intuitively, the time stamp is provided as an auxiliary input to help our network distinguish similar poses at different frames, while the VAE property can push the latent space to be uninformative, thereby encouraging the network to mainly rely on pose conditions when inferring node-related variables. With all of these building blocks, our network can be trained in an end-to-end manner, eventually producing a realistic dynamic human avatar.

Overall, our proposed method offers the new ability to automatically create an animatable human character with general, dynamic garments. This is achieved by using only RGB videos, without any pre-scanning efforts. Compared to methods that heavily depend on the topology of a naked human body template, our approach is powerful yet general in terms of both appearance learning and motion modeling, and able to generate realistic dynamic details. To the best of our knowledge, our method is the first one that demonstrates automatic human avatar creation for dresses. Experiments prove that our method outperforms state-of-the-art approaches qualitatively and quantitatively.

2. Related Work

Image-based 3D Human Reconstruction. Three-dimensional human character reconstruction is traditionally the very first step towards human avatar modeling. Previous studies focused on using multi-view images [38, 69, 73, 75, 76] or RGB(D) image sequences [3, 4, 7, 12, 13, 21, 23, 79, 80, 82–84, 88] for human model reconstruction. Extremely high-quality reconstruction results have also been demonstrated with tens or even hundreds of cameras [10]. In order to reduce the difficulty in system setup, human model reconstruction from sparse camera views has been investigated by using neural networks for learning silhouette cues [19, 48] and stereo cues [26]. More recently, various approaches were proposed to reconstruct a 3D human model from a single-view RGB images [5, 14, 25, 27, 61, 62, 74, 89, 90]. For example, PIFu [61] and PIFuHD [62] proposed to regress a deep implicit function using pixel-aligned image features and is able to reconstruct high-resolution results. ARCH [27] and ARCH++ [25] proposed to reconstruct the 3D human model in a canonical pose in order to support animation. Although demonstrating plausible results, these methods rely on large scale dataset of 3D human scans to train the model, and suffer from reconstruction errors and weak generalization capability. In contrast, our method bypasses the reconstruction step and directly learns an animatable avatar from RGB videos.

Neural Scene Representations and Rendering. Representing objects or scenes implicitly with neural networks, is becoming more and more popular for its compactness

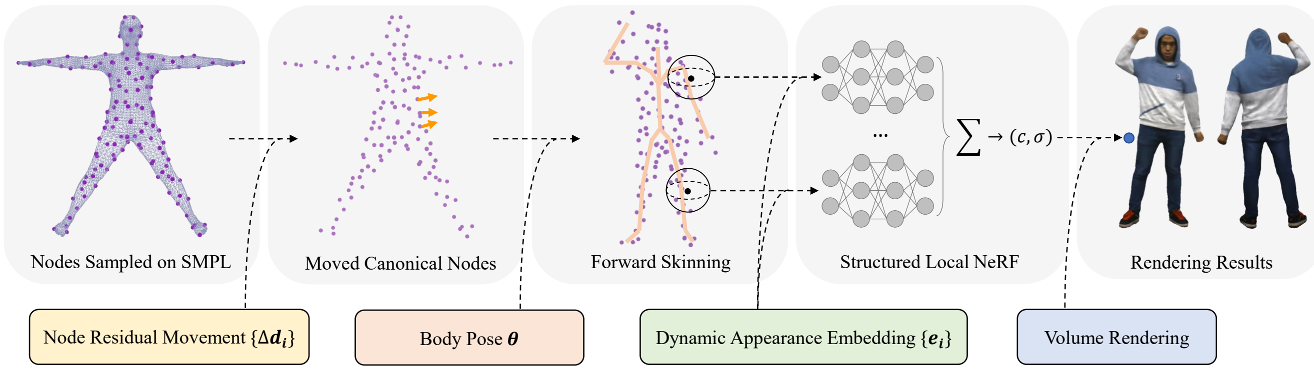

Figure 2. **Illustration of our clothed human representation.** In our proposed method, we represent the dynamic appearance of a clothed human character using structured local radiance fields attached to pre-defined nodes on the SMPL model. The garment deformations are then modeled in a coarse-to-fine manner with three set of variables, including the body poses as the coarsest level, the node residual translations as the middle level and the dynamic detail embeddings of the local radiance fields as the finest level.

and strong representation power. Pioneer studies proposed to learn an implicit function where the shapes are embedded into the iso-surface of network output [8, 9, 11, 18, 45, 51, 87]. Another line of work on implicit representation aimed at learning scene representations for novel view synthesis from posed 2D images. They represent static scenes using voxel grids of high-dimensional features [66], continuous learnable function [67] or neural radiance fields (NeRF) [46]. NeRF, in particular, shows strong capability of modeling view-dependent effects and thus attracts much attention [17, 34, 39, 44, 59, 81, 85]. It is later extended for dynamic scenes through deformation learning [15,16,32,33,52,57,64,72,77]. Human motions are usually much more challenging to learn using neural networks, and several works [30, 49, 55] incorporated prior from a statistical body template to tackle this difficulty. Note that most of these works can only playback the dynamic sequence that the networks are trained on, while our work aims at animation, which is a much harder task because the method has to generalize to new poses.

Animatable Human Avatars. In the last decade, many efforts have been made for achieving expressive and animatable 3D models for human avatars. To facilitate geometric learning, several statistical parametric templates are developed for face [31], hands [47, 60] and minimally clothed body [28,40,50,53]. To acquire animatable characters wearing casual clothes, traditional pipelines mostly reconstruct a subject-specific mesh template in advance, and then generate its motions using physics simulation [20, 70], deformation space modeling [28], or deep learning [6, 22, 78]. The reliance on pre-scanning efforts can be eliminated via deforming a general body template, and several works proposed to directly learn this deformation from geometric data [41–43, 56] or RGB videos [2–5]. The texture map and the rasterization step in those methods are later replaced with neural texture maps and image decoders in order to achieve photo-realistic rendering [36, 37, 58, 65]. Recently, neural scene representations and rendering techniques are adopted for higher-fidelity results [35, 54, 55]. However, state-of-the-art methods only demonstrate results of tightly-fitting garments, while our method is more general in terms of clothes topology and deformation.

3. Representation

Our goal is to learn an animatable virtual characters directly from RGB videos and to support loose clothes like skirts and dresses without pre-scanning a template. To this end, we propose a new representation that has a strong capability of modeling the shape, appearance and dynamic deformations of clothed humans. At its core is a set of *structured local radiance fields*, each of which models the dynamic appearance inside a local space while moving according to the body poses as well as the cloth deformations. To be more specific, we first pre-define N nodes on the SMPL model via farthest point sampling. Their coordinates on the canonical SMPL surface are denoted with $\{\bar{\boldsymbol{n}}_i\}_{i=1}^N$. Since the nodes are sampled from the SMPL model, each of them has an associated skinning weight vector $\boldsymbol{\omega}_i \in \mathbb{R}^J$, where J is the number of body joints. Given a pose vector $\boldsymbol{\theta}^{(t)}$ at time stamp t, we can transform node i to the posed space using linear blending skinning (LBS):

$$\mathbf{T}_i^{(t)} = \sum \omega_{i,j} \mathbf{M}_j(\boldsymbol{\theta}^{(t)}), \tag{1}$$

$$\boldsymbol{n}_i^{(t)} = \mathbf{T}_i^{(t)} \bar{\boldsymbol{n}}_i, \tag{2}$$

where $\mathbf{M}_j(\boldsymbol{\theta}^{(t)}) \in SE(3)$ is the rigid transformation of the j-th body joints and $\omega_{i,j}$ is the j-th entry of $\boldsymbol{\omega}_i$.

In Eqn. (2), the nodes strictly follow the motion of the body surface. In order to handle the non-rigid deformations of clothes, we allow the nodes to shift independently. Mathematically, we assign a time-varying residual translation $\Delta \boldsymbol{n}_i^{(t)}$ to node i in the canonical space, and modify

Eqn. (2) into:

$$\boldsymbol{n}_i^{(t)} = \mathbf{T}_i^{(t)} \left(\bar{\boldsymbol{n}}_i + \Delta \boldsymbol{n}_i^{(t)} \right). \tag{3}$$

Finally, we construct a local radiance field over the influence of each node, with a function $\mathcal{F}_i$ represented by a tiny MLP. This MLP takes as input a coordinate in the local space of node i and outputs a high-dimensional feature vector. To model the fine-grain dynamic details that cannot be represented by node translations, we condition the local radiance field on a dynamic detail embedding $\boldsymbol{e}_i^{(t)}$. Formally, given any point $\boldsymbol{p} \in \mathbb{R}^3$ in the posed space at frame t, we first calculate its coordinate in the local space of node i as:

$$\boldsymbol{p}_i = \left(\mathbf{T}_i^{(t)} \right)^{-1} \boldsymbol{p} - \left(\bar{\boldsymbol{n}}_i + \Delta \boldsymbol{n}_i^{(t)} \right). \tag{4}$$

After that, we feed it into the local radiance network $\mathcal{F}_i$ and blend the feature vectors produced by all local MLPs:

$$\boldsymbol{f} = \frac{\sum w_i \mathcal{F}_i(\boldsymbol{p}_i; \boldsymbol{e}_i^{(t)})}{\sum w_i}, \tag{5}$$

where w_i is the blending weight defined as

$$w_i = \max\{\exp(-\|\boldsymbol{p} - \boldsymbol{n}_i^{(t)}\|_2^2 / 2\sigma^2) - \epsilon, 0\}, \tag{6}$$

and ϵ is a hyperparameter controlling the influence radius of the nodes. This blended feature $\boldsymbol{f}$ is fed into two additional MLPs, $\mathcal{G}(\cdot)$ and $\mathcal{H}(\cdot)$, to compute the color & density of $\boldsymbol{p}$:

$$\text{Color}(\boldsymbol{p}) = \mathcal{G}(\boldsymbol{f}, \boldsymbol{v}), \quad \text{Density}(\boldsymbol{p}) = \mathcal{H}(\boldsymbol{f}), \tag{7}$$

where $\boldsymbol{v} \in \mathbb{R}^3$ is the viewing direction [46].

Overall, the dynamic appearance of a clothed character is parameterized in a coarse-to-fine fashion with three sets of variables: body poses $\{\boldsymbol{\theta}^{(t)}\}$, node residual translations $\{\Delta \boldsymbol{n}_i^{(t)}\}$ and dynamic detail embeddings $\{\boldsymbol{e}_i^{(t)}\}$. With the radiance field determined by these variables and the networks (*i.e.*, $\mathcal{F}_1, \mathcal{F}_2, ..., \mathcal{F}_N$, $\mathcal{G}$ and $\mathcal{H}$), we can shoot rays and render images via volume rendering as in [46]. An illustration of our representation is presented in Fig. 2.

Discussion. Compared to state-of-the-art methods, our representation has two advantages:

- Our method has expressive representation power in terms of both the motion and the topology. Although the nodes in our representation are sampled from the SMPL model, our method is not restricted by it. Instead, our method allows more degrees of freedom for motion and geometry modeling, enabling avatar creation for different cloth topologies, which is a significant departure from the existing works [35, 54, 58, 65].
- Our method does not explicitly define a global canonical field and consequently avoids the need for "backward skinning" during training. Backward skinning is used to transform the points in the posed space to a global canonical space, and has been the basis of previous methods [35, 54, 63]. Even so, we argue that this operation is ambiguous, especially for the points around contacting body parts. In contrast, our approach computes the radiance of any point in the local space, thus resolving the ambiguity issue.

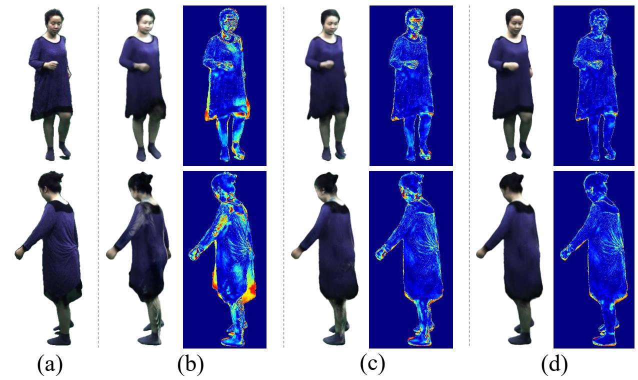

Figure 3. **Visualization of the effect of node-related variables.** (a) Ground-truth reference. (b) Rendering results without node residual translation and dynamic detail embeddings. (c) Results without dynamic detail embeddings. (d) Results with full set of variables. See Sec. 5.3 for details.

4. Method

Having elaborated on the proposed representation, we turn to network learning in this section. Specifically, we need to determine the aforementioned variables alongside with the weights of the radiance networks for a training image sequence $\boldsymbol{I}_t, t = 1, 2, ..., T$. The images can be captured from a multi-view system or a monocular one. In order to synthesize images for new poses, we also have to compute the node residual translations and the dynamic detail embeddings corresponding to those poses. We assume access to the body poses of the training images (*i.e.*, $\boldsymbol{\theta}^{(t)}, t = 1, 2, ..., T$), which can be estimated using markerless MoCap tools such as [1, 86]. The node residual translations and the detail embeddings are referred to as "node-related variables" in the following context.

4.1. Network Architecture

To obtain the node-related variables for the training frames and ensure generalization during animation, we design a simple conditional variational auto-encoders (cVAE) [68] as an auxiliary network for each node. Each auxiliary network consists of an encoder and a decoder, both implemented with tiny MLPs. Following the practice of SCANimate [63], the condition variable of this cVAE is the pose vector multiplied by the skinning weight and an attention map:

$$\boldsymbol{\theta}_i^{(t)} = (\mathbf{W} \cdot \boldsymbol{\omega}_i) \circ \boldsymbol{\theta}^{(t)}, \tag{8}$$

where $\mathbf{W}$ is the weight map that converts the skinning weights into pose attention weights as in [63] and $\circ$ denotes element-wise product. During training, the encoder takes the time stamp t as input and $\boldsymbol{\theta}_i^{(t)}$ as condition, and produces parameters of a Gaussian distribution, from which a latent code $\boldsymbol{z}_i^{(t)}$ is sampled:

$$\boldsymbol{\mu}_i^{(t)}, \boldsymbol{\sigma}_i^{(t)} \leftarrow \mathcal{E}(t, \boldsymbol{\theta}_i^{(t)}),\ \boldsymbol{z}_i^{(t)} \sim \mathcal{N}(\boldsymbol{\mu}_i^{(t)}, \boldsymbol{\sigma}_i^{(t)}), \quad (9)$$

Conditioned on the body pose, the latent code is then decoded into the node residual translation and the dynamic detail embedding:

$$\Delta \boldsymbol{n}_i^{(t)}, \boldsymbol{e}_i^{(t)} \leftarrow \mathcal{D}(\boldsymbol{z}_i^{(t)}, \boldsymbol{\theta}_i^{(t)}), \quad (10)$$

which are later used in Eqn. (4) and Eqn. (5), respectively.

In this network, the time instant is used to distinguish similar poses at different time instants, thereby avoiding the one-to-many mapping issue. With the KL-divergence loss in cVAE, there is a preference to let the decoder to mainly rely on the pose condition for prediction, and the time input only provides information necessary for good reconstruction. In our implementation, we augment the time stamp and the coordinates with Fourier encoding before feeding them into MLPs [46] . Fig. 4 illustrates the data flow in our network during training. Once the training is done, we can render the model for either training frames or novel poses. To render the training sequence, we use the full network and set $\boldsymbol{z}_i^{(t)} = \boldsymbol{\mu}_i^{(t)}$ in Eqn. (9) to eliminate randomness. When unseen poses are given, the encoder half of the cVAE will be omitted and $\boldsymbol{z}_i^{(t)}$ will be set to zeros.

4.2. Training Loss

Our network can be trained in an end-to-end manner. The training loss is composed of four components, including a reconstruction loss, a node translation regularization, an embedding regularization, and a KL-divergence loss:

$$\mathcal{L} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{trans}\mathcal{L}_{trans} + \lambda_{ebd}\mathcal{L}_{ebd} + \lambda_{KL}\mathcal{L}_{KL}. \quad (11)$$

Below we discuss them in details. For ease of notation, we drop the superscript $^{(t)}$ of all variables in this subsection.

Reconstruction Loss $\mathcal{L}_{rec}$ measures the mean squared error between the rendered and true pixel colors:

$$\mathcal{L}_{rec} = \sum_{\boldsymbol{r} \in \mathcal{R}} \left\| \boldsymbol{C}\left(\boldsymbol{r} | \boldsymbol{\theta}, \{\Delta \boldsymbol{n}_i\}, \{\boldsymbol{e}_i\}\right) - \hat{\boldsymbol{C}}_{\boldsymbol{r}} \right\|_2^2, \quad (12)$$

where $\mathcal{R}$ is the set of rays in each batch, $\hat{\boldsymbol{C}}_{\boldsymbol{r}}$ is the ground-truth pixel color, $\boldsymbol{C}(\cdot|\boldsymbol{\theta}, \{\Delta \boldsymbol{n}_i\}, \{\boldsymbol{e}_i\})$ is the volume rendering function with the representation defined in Sec. 3.

Node Translation Regularization $\mathcal{L}_{trans}$ simply constrains the position change of each nodes in order to stabilize training:

$$\mathcal{L}_{trans} = \sum_i \|\Delta \boldsymbol{n}_i\|_2^2. \quad (13)$$

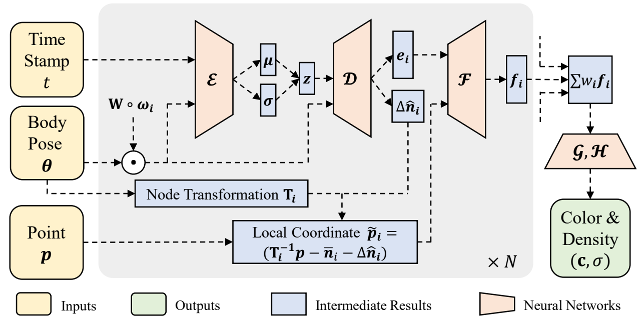

Figure 4. **Illustration of the data flow in our network.** The time stamp and body pose feature are first passed through the cVAEs, which produces the node residual translations and dynamic detail embeddings of the local radiance fields. For a point in the posed space, we calculate its local coordinate in each local field, and then query its feature. Finally, all features are blended and decoded into the color and density values.

Embedding Regularization $\mathcal{L}_{ebd}$ penalize large magnitudes of the dynamic detail embeddings:

$$\mathcal{L}_{ebd} = \sum_i \|\boldsymbol{e}_i\|_2^2. \quad (14)$$

A similar loss is also used in [51]; here we utilize it to encourage the embeddings to encode only the information that cannot be represented by node position.

KL-divergence Loss $\mathcal{L}_{KL}$ is a standard VAE KL-divergence penalty [29]:

$$\mathcal{L}_{KL} = \sum_i \mathrm{KL}\left(\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i) \,\|\, \mathcal{N}(\mathbf{0}, \mathbf{I})\right). \quad (15)$$

Implementation Details The local radiance networks and cVAEs in our architecture are implemented with parallel tiny MLPs in the form of group 1D convolution. To accelerate training and inference, we exploit the fact that, for any point in the posed space, only a small portion of nodes have influence on its color and density value. We use Adam optimizer to train our models. Training the whole models takes about 25 hours on one NVIDIA 3090 GPU with 500k iterations, while rendering an color image with resolution of 512×512 typically takes 5 seconds on one NVIDIA 3080TI GPU. Please refer to the *Supp.Mat.* for more details.

5. Experiments

Dataset and Metrics. For evaluation and comparison with baseline methods, we mainly use the following dataset: (1) Two dress sequences from [22], which are captured using 100 cameras but we manually select 20 views among them for computational efficiency; (2) One sweater sequences from [24] captured with 10 cameras; (3) Two sequences from ZJU-MoCap [55] captured with 23 cameras; and (4) three multi-view sequences collected by ourselves with 24 cameras[1]. For quantitative evaluation, we use two standard

[1] Data collection and disclosure have been consented by the volunteers.

Figure 5. **Example results of our method.** We train our network on various datasets and show the novel pose synthesis results.

metrics: peak signal-tonoise ratio (PSNR) and structural similarity index (SSIM). More details about data collection and preprocessing can be found in the *Supp.Mat.*.

5.1. Results

We train our model for each individual subjects, and present some example animation results in Fig. 1 and Fig. 5. The results cover various body poses and different cloth styles. As shown in these figures, our method not only gracefully tackles different cloth types, but also generates realistic dynamic wrinkles. Please see our supplemental video for more visualization.

Although we mainly use multi-view videos for evaluation, our method is also able to learn an avatar from single-view input. Fig. 6 demonstrates the results of our method on the PeopleSnapshot dataset [4], which captures performers rotating 360 degrees in an A-pose with a monocular camera. As shown in the figure, our method can also work well with such extremely simple input, further proving its generalization capability.

5.2. Comparison

We mainly compare our method with Animatable NeRF [54] and Neural Body [55]. We omit other related methods since they have been compared in [54].

We first compare with Animatable NeRF [54] on the dataset of [22] and our own data. We split each video into training frames and testing ones, train the networks using the training frames from all views, and test the animation quality using the testing frames. Qualitative results are presented in Fig. 7. Compared to [54], our method can produce more appearance details, and generate the non-rigid mo-

Figure 6. **Our results on PeopleSnapshot dataset.** Given a monocular video recording a person rotating in an A-pose (top), our method is able to create a human avatar that supports novel pose generation and free view synthesis (bottom).

Figure 7. **Comparison against Animatable Nerf** [54] on novel pose synthesis.

tions of dress hems. The numeric results in Tab. 1 also prove that our method can achieve higher-quality results than [54].

To conduct a fair comparison with Neural Body [55], we use their dataset and follow the same protocal in their paper. In this comparison, we train our network using only 300 image frames from four views, as done in [55]. We evaluate the quality of novel view synthesis for training frames and unseen body poses. The results in Tab. 2 shows that our model achieves higher accuracy than [55] in both metrics. In fact, our method performs better not only in learning appearance details like the logo, but also in generalizing to unseen poses, as shown in Fig. 8. We also report the numeric results of Animatable NeRF [54] in Tab. 2 for completeness.

Table 1. Quantitative comparison with Animatable NeRF [54] in terms of novel pose synthesis.

	PSNR (↑)		SSIM (↑)	
Case \ Method	[54]	Ours	[54]	Ours
Hoody	22.43	**24.94**	0.893	**0.928**
Jacket	24.30	**25.24**	0.909	**0.927**
Dress1	19.52	**23.43**	0.848	**0.891**
Dress2	20.49	**22.19**	0.877	**0.900**

Table 2. Quantitative comparison with Neural Body [55] and Animatable NeRF [54] on ZJU-MoCap dataset.

		PSNR (↑)			SSIM (↑)		
ID	Pose Type	[55]	[54]	Ours	[55]	[54]	Ours
387	Seen	25.79	24.38	**28.32**	0.928	0.903	**0.953**
	Unseen	21.60	21.29	**23.61**	0.870	0.860	**0.905**
392	Seen	29.44	27.43	**30.79**	0.946	0.919	**0.958**
	Unseen	25.76	24.59	**26.74**	0.909	0.889	**0.927**

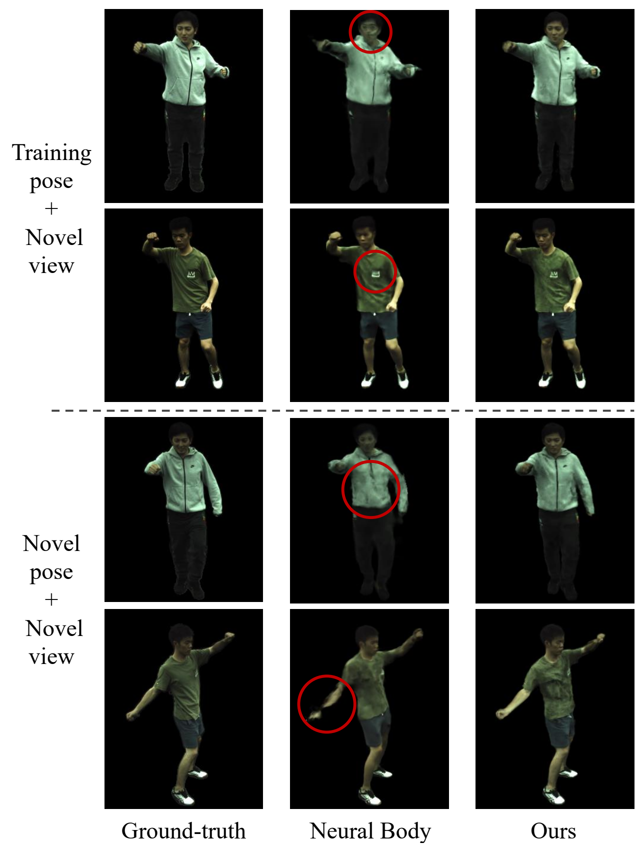

Figure 8. **Comparison against Neural Body** [55] in terms of both novel view synthesis and pose generation. Zoom in for better view.

Figure 9. **Evaluation of our cVAE design.** We replace the cVAE with a determinstic regression network, and compare the reconstruction results of training frames. (a,d) Ground-truth. (b,e) Results by the deterministic baseline. (c,f) Our results.

5.3. Ablation Study

In this subsection, we conduct three qualitative ablation experiments on the main components of our method design. We present the quantitative results as well as some additional experiments in the *Supp.Mat.*.

Node-related variables. To understand the effect of the node-related variables in our method, we take the trained model for a dress sequence and conduct experiment on it. Specifically, we render the images of training poses under three circumstances, *i.e.*, 1) without node residual translations or dynamic detail embeddings, 2) with node residual translations but without dynamic detail embeddings, and 3) with both node translations and detail embeddings. The results are shown in Fig. 3. As the figure shows, when the node residual translations and the dynamic detail embeddings are both disabled, the model only recovers the articulated motions and fails to render the correct shape of the moving character. With solely the node residual translation enabled, the non-rigid deformation of the dress hem can be recovered, but the shading on the facial area is not consistent with the image evidence. Only with both the node residual translation and the dynamic detail embeddings enabled can all appearance details be faithfully reconstructed.

cVAE. We evaluate our choice of cVAE-based architecture by replacing it with a deterministic network that directly regresses the node-related variables from body poses. This baseline network is trained under the same setting as our proposed model. We render the images for training frames in order to compare the performance of data fitting, and the results are presented in Fig. 9. Not surprisingly, naively learning a mapping from pose parameters to the node-related variables, without specifically account for the potential one-to-many mapping problem, will produce averaged appearance and fail to recover the dynamic garment wrinkles even for training images. In contrast, our method can fit to training data much better than the baseline method, consequently enabling realistic animation and rendering.

Time stamp input. There exist other options that can be used as the cVAE input for resolving the one-to-many mapping problem. For instance, we can use learnable per-frame latent embeddings. The motivation behind our choice of time stamp is that, the low-frequency bias in MLPs can

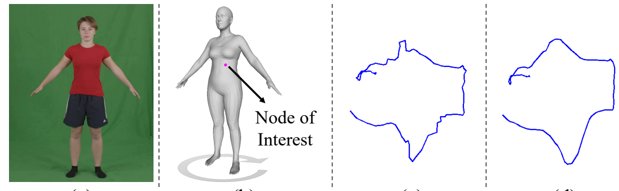

Figure 10. **Evaluation of the time instant input.** We replace the time stamp input with learnable per-frame latent codes and compare the trajectory of nodes. (a) Training video. (b) The node of which we visualize the trajectory. (c) Node trajectory using learnable latent codes. (d) Node trajectory using the proposed method.

ensure temporal smoothness of the node-related variables, especially for node residual translations. In this way, we avoid the need for an additional loss of temporal smoothness. To validate this motivation, we conduct an ablation study where we replace the time stamp input with learnable latent embeddings. Then we compare the node trajectories as in Fig. 10. As shown by the results, without explicitly constraining the temporal smoothness, the baseline method learns noisy node motions, while the trajectory of our method is much more smooth and physically plausible.

6. Discussion

Conclusion. We introduced a novel method that uses structured local radiance fields for generation of controllable clothed human avatars. It has expressive representation power for both appearance and motion, as we leverage the advantages of neural scene representation while explicitly accounting for the motion hierarchy of clothes. Compared to existing methods, ours can handle more general cloth styles and generate realistic dynamic details.

Limitation. The performance of our method depends on the pose variance in the training data, and our method may fail to generate plausible results when the animation poses starkly differ from the training poses; see Supp.Mat. for an example. In addition, the dynamic deformations and wrinkle changes of garments involve complex physics processes, which may be beyond the representation capability of our model. Finally, our method assumes accurate body pose estimation for the training images; that is why we mainly conduct experiments on multi-view dataset. For monocular videos, erroneous pose estimation caused by ambiguity may eventually lead to rendering artifacts.

Potential Social Impact. Our method enables automatic creation of a digital twin of any person. It can be combined with existing Deep Fake algorithms to generate fake videos through character animation and reenactment, which need to be addressed carefully before deploying the technology.

Acknowledgement. This paper is sponsored by National Key R&D Program of China (2021ZD0113503) and the NSFC No. 62125107 and No. 62171255.

References

[1] https://github.com/zju3dv/EasyMocap. 4

[2] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single RGB camera. In *CVPR*, 2019. 1, 3

[3] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *3DV*, 2018. 1, 2, 3

[4] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *CVPR*, 2018. 1, 2, 3, 6

[5] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *ICCV*, 2019. 2, 3

[6] Timur Bagautdinov, Chenglei Wu, Tomas Simon, Fabián Prada, Takaaki Shiratori, Shih-En Wei, Weipeng Xu, Yaser Sheikh, and Jason Saragih. Driving-signal aware full-body avatars. *ACM TOG*, 40(4), July 2021. 1, 2, 3

[7] Federica Bogo, Michael J Black, Matthew Loper, and Javier Romero. Detailed full-body reconstructions of moving people from monocular rgb-d sequences. In *ICCV*, 2015. 2

[8] Aljaz Bozic, Pablo R. Palafox, Michael Zollhöfer, Justus Thies, Angela Dai, and Matthias Nießner. Neural deformation graphs for globally-consistent non-rigid reconstruction. In *CVPR*, 2021. 3

[9] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *CVPR*, 2019. 3

[10] Alvaro Collet, Ming Chuang, Pat Sweeney, Don Gillett, Dennis Evseev, David Calabrese, Hugues Hoppe, Adam Kirk, and Steve Sullivan. High-quality streamable free-viewpoint video. *ACM TOG*, 34(4):69, 2015. 2

[11] Boyang Deng, JP Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Nasa: Neural articulated shape approximation. In *ECCV*, 2020. 3

[12] Mingsong Dou, Philip L. Davidson, Sean Ryan Fanello, Sameh Khamis, Adarsh Kowdle, Christoph Rhemann, Vladimir Tankovich, and Shahram Izadi. Motion2fusion: real-time volumetric performance capture. *ACM TOG*, 36(6):246:1–246:16, 2017. 2

[13] Mingsong Dou, Sameh Khamis, Yury Degtyarev, Philip Davidson, Sean Ryan Fanello, Adarsh Kowdle, Sergio Orts Escolano, Christoph Rhemann, David Kim, Jonathan Taylor, et al. Fusion4d: Real-time performance capture of challenging scenes. *ACM TOG*, 35(4):1–13, 2016. 2

[14] Valentin Gabeur, Jean-Sébastien Franco, Xavier Martin, Cordelia Schmid, and Gregory Rogez. Moulding humans: Non-parametric 3d human shape estimation from single images. In *ICCV*, 2019. 2

[15] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *CVPR*, 2021. 3

[16] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *ICCV*, 2021. 2, 3

[17] Stephan J. Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien P. C. Valentin. Fastnerf: High-fidelity neural rendering at 200fps. In *ICCV*, 2021. 3

[18] Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas A. Funkhouser. Local deep implicit functions for 3d shape. In *CVPR*, 2020. 3

[19] Andrew Gilbert, Marco Volino, John Collomosse, and Adrian Hilton. Volumetric performance capture from minimal camera viewpoints. In *ECCV*, 2018. 2

[20] P. Guan, L. Reiss, D. Hirshberg, A. Weiss, and M. J. Black. DRAPE: DRessing Any PErson. *ACM TOG (Proc. SIGGRAPH)*, 31(4):35:1–35:10, July 2012. 3

[21] Kaiwen Guo, Feng Xu, Tao Yu, Xiaoyang Liu, Qionghai Dai, and Yebin Liu. Real-time geometry, albedo and motion reconstruction using a single rgbd camera. *ACM TOG*, 36(3):32:1–32:13, 2017. 2

[22] Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Real-time deep dynamic characters. *ACM TOG*, 40(4), aug 2021. 1, 3, 5, 6

[23] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Livecap: Real-time human performance capture from monocular video. *ACM TOG*, 38(2):1–17, 2019. 2

[24] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *CVPR*, 2020. 5

[25] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *ICCV*, 2021. 2

[26] Zeng Huang, Tianye Li, Weikai Chen, Yajie Zhao, Jun Xing, Chloe LeGendre, Linjie Luo, Chongyang Ma, and Hao Li. Deep volumetric video from very sparse multi-view performance capture. In *ECCV*, 2018. 2

[27] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *CVPR*, 2020. 2

[28] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *CVPR*, 2018. 3

[29] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 5

[30] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. In *Advances in Neural Information Processing Systems*, 2021. 3

[31] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM TOG (SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 3

[32] Tianye Li, Mira Slavcheva, Michael Zollhöfer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, and Zhaoyang Lv. Neural 3d video synthesis. *CoRR*, abs/2103.02597, 2021. 3

[33] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *CVPR*, 2021. 3

[34] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *NeurIPS*, 2020. 3

[35] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM TOG (ACM SIGGRAPH Asia)*, 2021. 1, 2, 3, 4

[36] Lingjie Liu, Weipeng Xu, Marc Habermann, Michael Zollhöfer, Florian Bernard, Hyeongwoo Kim, Wenping Wang, and Christian Theobalt. Neural human video rendering by learning dynamic textures and rendering-to-video translation. *IEEE TVCG*, PP:1–1, 05 2020. 1, 3

[37] Lingjie Liu, Weipeng Xu, Michael Zollhöfer, Hyeongwoo Kim, Florian Bernard, Marc Habermann, Wenping Wang, and Christian Theobalt. Neural rendering and reenactment of human actor videos. *ACM TOG*, 38(5):139:1–139:14, 2019. 1, 3

[38] Yebin Liu, Qionghai Dai, and Wenli Xu. A point-cloud-based multiview stereo algorithm for free-viewpoint video. *IEEE TVCG*, 16(3):407–418, 2009. 2

[39] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhöfer, Yaser Sheikh, and Jason M. Saragih. Mixture of volumetric primitives for efficient neural rendering. *ACM TOG*, 40(4):59:1–59:13, 2021. 3

[40] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM TOG*, 34(6):1–16, 2015. 1, 3

[41] Qianli Ma, Shunsuke Saito, Jinlong Yang, Siyu Tang, and Michael J. Black. SCALE: Modeling clothed humans with a surface codec of articulated local elements. In *CVPR*, 2021. 3

[42] Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J. Black. Learning to Dress 3D People in Generative Clothing. In *CVPR*, 2020. 3

[43] Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J. Black. The power of points for modeling humans in clothing. In *ICCV*, 2021. 3

[44] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *CVPR*, 2021. 3

[45] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, 2019. 3

[46] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2, 3, 4, 5

[47] Gyeongsik Moon, Takaaki Shiratori, and Kyoung Mu Lee. Deephandmesh: A weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling. In *ECCV*, 2020. 3

[48] Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. Siclope: Silhouette-based clothed people. In *CVPR*, 2019. 2

[49] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *ICCV*, 2021. 1, 2, 3

[50] Ahmed A A Osman, Timo Bolkart, and Michael J. Black. STAR: A sparse trained articulated human body regressor. In *ECCV*, 2020. 3

[51] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019. 3, 5

[52] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *ICCV*, 2021. 3

[53] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 3

[54] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *ICCV*, 2021. 1, 2, 3, 4, 6, 7

[55] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021. 3, 5, 6, 7

[56] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM TOG*, 36(4):1–15, 2017. 3

[57] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *CVPR*, 2021. 2, 3

[58] Amit Raj, Julian Tanke, James Hays, Minh Vo, Carsten Stoll, and Christoph Lassner. ANR: articulated neural rendering for virtual avatars. In *CVPR*, 2021. 1, 3, 4

[59] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *ICCV*, 2021. 3

[60] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM TOG*, 36(6), Nov. 2017. 3

[61] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, 2019. 2

[62] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*, 2020. 2

[63] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J. Black. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *CVPR*, 2021. 4, 5

[64] Ruizhi Shao, Hongwen Zhang, He Zhang, Mingjia Chen, Yanpei Cao, Tao Yu, and Yebin Liu. Doublefield: Bridging the neural surface and radiance fields for high-fidelity human reconstruction and rendering. In *CVPR*, 2022. 3

[65] Aliaksandra Shysheya, Egor Zakharov, Kara-Ali Aliev, Renat Bashirov, Egor Burkov, Karim Iskakov, Aleksei

Ivakhnenko, Yury Malkov, Igor Pasechnik, Dmitry Ulyanov, Alexander Vakhitov, and Victor Lempitsky. Textured neural avatars. In *CVPR*, 2019. 1, 3, 4

[66] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhöfer. Deepvoxels: Learning persistent 3d feature embeddings. In *CVPR*, 2019. 3

[67] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *NeurIPS*, 2019. 3

[68] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *NIPS*, 2015. 2, 4

[69] Jonathan Starck and Adrian Hilton. Surface capture for performance-based animation. *IEEE Computer Graphics and Applications*, 27(3):21–31, 2007. 2

[70] Carsten Stoll, Juergen Gall, Edilson de Aguiar, Sebastian Thrun, and Christian Theobalt. Video-based reconstruction of animatable human characters. *ACM TOG*, 29(6), Dec. 2010. 3

[71] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *NeurIPS*, 2020. 2

[72] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a deforming scene from monocular video. *CoRR*, abs/2012.12247, 2020. 3

[73] Daniel Vlasic, Pieter Peers, Ilya Baran, Paul E. Debevec, Jovan Popovic, Szymon Rusinkiewicz, and Wojciech Matusik. Dynamic shape capture using multi-view photometric stereo. *ACM TOG*, 28(5):174:1–174:11, 2009. 2

[74] Lizhen Wang, Xiaochen Zhao, Tao Yu, Songtao Wang, and Yebin Liu. Normalgan: Learning detailed 3d human from a single rgb-d image. In *ECCV*, 2020. 2

[75] Michael Waschbüsch, Stephan Würmlin, Daniel Cotting, Filip Sadlo, and Markus H. Gross. Scalable 3d video of dynamic scenes. *The Visual Computer*, 21(8-10):629–638, 2005. 2

[76] Chenglei Wu, Kiran Varanasi, Yebin Liu, Hans-Peter Seidel, and Christian Theobalt. Shading-based dynamic shape refinement from multi-view video under general illumination. In *ICCV*, 2011. 2

[77] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *CVPR*, 2021. 2, 3

[78] Donglai Xiang, Fabian Prada, Timur Bagautdinov, Weipeng Xu, Yuan Dong, He Wen, Jessica Hodgins, and Chenglei Wu. Modeling clothing as a separate layer for an animatable human avatar. *ACM TOG*, 40(6), 2021. 1, 3

[79] Donglai Xiang, Fabian Prada, Chenglei Wu, and Jessica K. Hodgins. Monoclothcap: Towards temporally coherent clothing capture from monocular RGB video. In *3DV*, 2020. 2

[80] Lan Xu, Zhuo Su, Lei Han, Tao Yu, Yebin Liu, and Lu Fang. Unstructuredfusion: Realtime 4d geometry and texture reconstruction using commercial RGBD cameras. *IEEE TPAMI*, 42(10):2508–2522, 2020. 2

[81] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *ICCV*, 2021. 3

[82] Tao Yu, Kaiwen Guo, Feng Xu, Yuan Dong, Zhaoqi Su, Jianhui Zhao, Jianguo Li, Qionghai Dai, and Yebin Liu. Bodyfusion: Real-time capture of human motion and surface geometry using a single depth camera. In *ICCV*, 2017. 2

[83] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *CVPR*, 2021. 2

[84] Tao Yu, Zerong Zheng, Kaiwen Guo, Jianhui Zhao, Qionghai Dai, Hao Li, Gerard Pons-Moll, and Yebin Liu. Doublefusion: Real-time capture of human performances with inner body shapes from a single depth sensor. In *CVPR*, 2018. 2

[85] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *CoRR*, abs/2010.07492, 2020. 3

[86] Yuxiang Zhang, Zhe Li, Liang An, Mengcheng Li, Tao Yu, and Yebin Liu. Light-weight multi-person total capture using sparse multi-view cameras. In *ICCV*, 2021. 4

[87] Zerong Zheng, Tao Yu, Qionghai Dai, and Yebin Liu. Deep implicit templates for 3d shape representation. In *CVPR*, 2021. 3

[88] Zerong Zheng, Tao Yu, Hao Li, Kaiwen Guo, Qionghai Dai, Lu Fang, and Yebin Liu. Hybridfusion: Real-time performance capture using a single depth sensor and sparse imus. In *ECCV*, 2018. 2

[89] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE TPAMI*, 2021. 2

[90] Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *CVPR*, 2019. 2