

Surface-Aligned Neural Radiance Fields for Controllable 3D Human Synthesis

Tianhan Xu*
 The University of Tokyo

tianhan.xu@mi.t.u-tokyo.ac.jp

Yasuhiro Fujita Eiichi Matsumoto
 Preferred Networks, Inc.

{fujita, matsumoto}@preferred.jp

Abstract

We propose a new method for reconstructing controllable implicit 3D human models from sparse multi-view RGB videos. Our method defines the neural scene representation on the mesh surface points and signed distances from the surface of a human body mesh. We identify an indistinguishability issue that arises when a point in 3D space is mapped to its nearest surface point on a mesh for learning surface-aligned neural scene representation. To address this issue, we propose projecting a point onto a mesh surface using a barycentric interpolation with modified vertex normals. Experiments with the ZJU-MoCap and Human3.6M datasets show that our approach achieves a higher quality in a novel-view and novel-pose synthesis than existing methods. We also demonstrate that our method easily supports the control of body shape and clothes. Project page: <https://pfnet-research.github.io/surface-aligned-nerf/>.

1. Introduction

Human body modeling is a long-studied topic for its wide range of real-world applications. In visual applications such as movies or games, which often require free-viewpoint rendering, it is common to expect 3D human models to have controllable properties such as pose, shape, and clothes. Because manually designing high-quality 3D human models is usually labor-intensive, increasing studies [1–3, 24, 27, 30, 35, 36] have proposed the reconstruction of 3D human models using only 2D observations. In this paper, we focus on the free-viewpoint 3D human synthesis with the above controllable properties from sparse multi-view RGB videos.

Early approaches [6, 42] deformed the pre-scanned template meshes with a skeleton for modeling a human shape and/or texture. Parametric 3D human models [4, 26, 34] have been proposed to reconstruct rough human meshes with pose and body shape estimation. Subsequent studies [1–3, 27] introduced more features, such as per-vertex deformation or texture, to express richer details. Such

parametric model-based approaches, despite having good controllable properties, show limitations in representing clothed humans, particularly when the actual shape differs significantly from the base parametric mesh estimation.

Neural radiance fields (NeRF) [28], a new form of 3D scene representation, has recently become the new baseline method in 3D reconstruction for its photorealistic rendering results of novel camera view. NeRF represents the scene as a continuous volumetric representation using a neural network to regress the color and density at a given query point from a given view direction. Several approaches [24, 30, 35, 36] have been proposed to incorporate knowledge from a statistical 3D human model and its pose estimation with NeRF. They differ in how a query point is transformed and represented. Deformation-based approaches [24, 35] use a deformation field to transform the query point from the observation space to a pose-independent canonical space and then build NeRF in the canonical space. Other approaches transform the query points into local coordinate systems [30] or to a latent code representation [36], with the help of human pose estimation.

In this paper, we propose a new approach for achieving a dynamic human reconstruction by combining parametric 3D body models such as SMPL [26] with NeRF. The basic idea of our approach is straightforward and simple: we propose building a NeRF on the mesh surface. We devise an algorithm to map a query point to a mesh surface point with a signed height, which can represent the local position of the point with respect to the mesh. Using the information of the surface point position and the signed height of a query point as input, we build a surface-aligned NeRF that is aligned with the mesh surface; thus, it can be easily deformed or controlled according to the base SMPL model.

Our approach has the following advantages: First, with the help of the devised mapping algorithm, our method does not rely on a learned deformation field, saving the number of learned parameters. Second, the models reconstructed using our method can be controlled directly by the SMPL parameters, that is, both poses and body shapes. Third, due to the surface-aligned property, our approach shows a better generalization ability for a novel human pose synthesis.

*Work done while the first author interned at Preferred Networks, Inc.

In summary, our contributions are as follows:

- We propose an algorithm that can injectively map a spatial point to a novel surface-aligned representation that consists of a projected surface point and a signed height to the mesh surface.
- We propose novel surface-aligned neural radiance fields using the proposed mapping, which can be easily controlled using the SMPL parameters. Compared to existing methods, our approach shows a better generalization performance on a novel view and novel pose synthesis while supporting manipulations such as changes to the body shape and clothes.

2. Related Work

Human body modeling. Early studies proposed parametric mesh models such as SCAPE [4] or SMPL [26] to model the shape of the human body. SMPL recovers the human mesh given the skeleton pose and body shape and is commonly used as the basis for controllable human modeling. However, one of the problems with the parametric model is that it can only model the naked human body. Subsequent studies proposed a per-vertex deformation based on the SMPL for better modeling of clothed-human details [1–3, 27]. Due to the strong expressiveness of the implicit 3D representation, recent studies proposed combining it with the SMPL to better capture the shape and appearance of a human body [5, 13]. In [44], the shape, pose, and skinning weights are represented with neural implicit functions to model a dynamic human body. In [8], a pose-conditioned implicit occupancy function is used to predict the shape of an articulated human body. There are also studies [29, 39, 40] that addressed reconstruction of clothed humans from a single image.

Neural radiance fields for a dynamic human body. Recently, neural radiance fields (NeRF) [28] has become a common building block for photorealistic novel view synthesis. Some studies have improved on the original NeRF to enable it to model dynamic scenes [9, 23, 32, 33, 37, 41]. The key idea of these methods is to introduce a deformation field that maps the observation space to a canonical space and builds a NeRF in the canonical space. However, optimizing both the deformation field and NeRF is an under-constrained problem that can cause implausible results or artifacts [32].

For modeling a dynamic human body, recent studies [24, 30, 35, 36] have proposed the use of prior knowledge of human pose and skinning weights of SMPL [26] to ease the learning of a deformation field. Animatable NeRF [35] uses the skinning weight of SMPL [26] to predict the neural blend shape fields. Neural Actor [24] predicts the texture map from the posed mesh and utilizes it as additional

information to help predict the deformation fields. However, learning a deformation field leads to a significant increase in network parameters, as well as the cost and difficulty of training and inference. Neural Body [36] uses the structured latent code anchored at the SMPL vertices to encode the pose information. However, this approach shows poor performance for a novel pose synthesis because the encoding process of the latent code strongly relies on the training pose. NARF [30] uses the human skeleton to transform spatial points into bone coordinates to learn a locally-defined NeRF, which has some conceptual similarity to our approach. However, because such bone coordinates have no surface information, they do not fully utilize a human body prior from SMPL nor support manipulations such as body shape control.

3. Method

We aim to reconstruct a 3D human model from sparse multi-view videos that can be photorealistically rendered with a fully controllable camera view, human pose, and body shape. We assume that, for each video frame, we have approximate human pose and shape information (*e.g.*, SMPL [26] parameters) and the foreground mask using off-the-shelf methods [10].

An overview of the proposed approach is shown in Fig. 1. We propose feeding a NeRF (Sec. 3.1) with a novel representation (Sec. 3.2) consisting of a surface point and signed height, which is calculated through the proposed dispersed projection (Sec. 3.3) given a query point. We call this variant of the NeRF a surface-aligned NeRF (Sec. 3.5).

3.1. Neural radiance fields revisited

Neural Radiance Fields (NeRF) [28] uses a neural network to model a scene in a continuous volumetric representation. A neural network f calculates the RGB color \mathbf{c} and the density σ at a given spatial position \mathbf{x} from a given viewpoint \mathbf{d} , which can be written as:

$$f : (\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma) \quad (1)$$

NeRF-based methods typically use position information $\mathbf{x} \in \mathbb{R}^3$ in a fixed three-dimensional Euclidean space (*e.g.*, the world coordinate). To model dynamic human bodies, recent studies have proposed some transformation or new representation of position information as the input of the NeRF. Specifically, NARF [30] transforms the position information of spatial points into each bone coordinate and uses all of them as input; Neural Body [36] uses a neural network to compute a latent code representing the position information relative to the body mesh and uses it as input.

3.2. Surface-aligned representation

We assume that the motion of the human body can be divided into the following three components according to

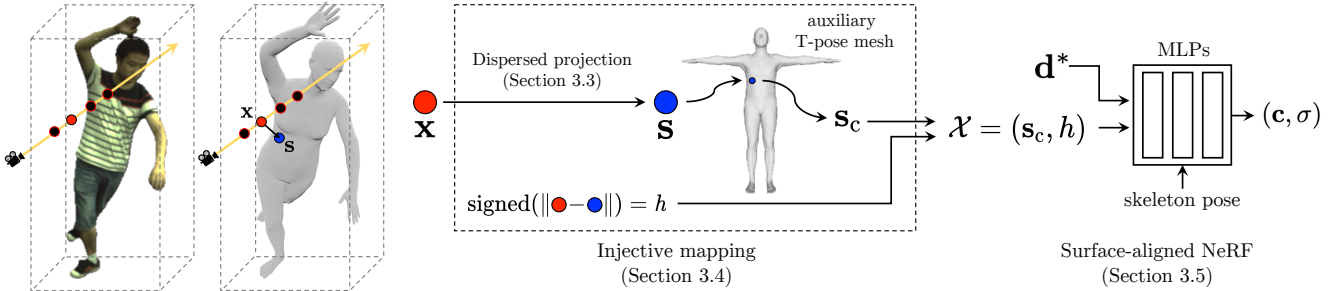


Figure 1. **An overview of our approach.** Given a query point $\mathbf{x} \in \mathbb{R}^3$, we use the proposed dispersed projection to project it onto a point $\mathbf{s} \in \mathbb{R}^3$ on the mesh surface to obtain a surface-aligned representation \mathcal{X} . The representation \mathcal{X} and the view direction \mathbf{d}^* are then input into the NeRF to compute the color \mathbf{c} and density σ of the query point \mathbf{x} .

the order of dominance: 1) surface-aligned components that follow the deformation of the body mesh, 2) pose-dependent components (e.g., clothes deformation caused by pose change), 3) the rest of the time-varying components (e.g., fluttering from the wind). We propose a new representation, called *surface-aligned representation*, for representing the positions of the query point $\mathbf{x} \in \mathbb{R}^3$ with respect to the body mesh, thus describing the most dominant mesh-following deformation components. Specifically, our proposed surface-aligned representation consists of two components: 1) The position information of surface point \mathbf{s} obtained by projecting \mathbf{x} onto the mesh surface. 2) The signed distance h between \mathbf{x} and the projection point \mathbf{s} , which represents the “height” of \mathbf{x} to the mesh surface. To represent \mathbf{s} in a pose-independent manner, we map it to the surface point on the shared T-pose body mesh. Specifically, for \mathbf{s} on the posed mesh surface, we compute the point with the same barycentric coordinate of the same triangle face on the T-pose mesh surface and use its spatial coordinates $\mathbf{s}_c = (x_c, y_c, z_c)$. The proposed surface-aligned representation can be written as:

$$\mathcal{X} = (\mathbf{s}_c, h) = (x_c, y_c, z_c, h) \in \mathbb{R}^4 \quad (2)$$

We map the spatial point \mathbf{x} to the surface-aligned representation \mathcal{X} and use it as input to the NeRF.

3.3. Dispersed projection

We propose a method of projecting a spatial point \mathbf{x} onto the mesh surface to obtain \mathbf{s} , called *dispersed projection*, which consists of two key components: *barycentric interpolated projection* and *vertex normal alignment*. We first introduce them separately and then show the detailed steps of the dispersed projection. Note that in the following we *only consider the case of $h > 0$ (a point outside the mesh)* because the idea is exactly the same as when points are inside (in that case, we can revert the normal direction). We assume that each vertex normals are from the inside of the mesh to the outside, that is, formally, the inner products are

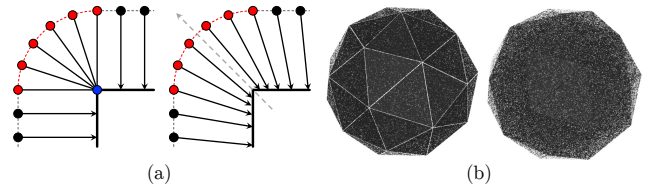


Figure 2. **Comparison of different projection methods.** (a) Toy example in 2D. We project the points of the same height onto the surface (thick line). Left: nearest point projection projects all the red points onto the vertex, thus obtaining the same surface-aligned representation. Right: proposed dispersed projection (here equals to the barycentric interpolated projection), in which all points can be projected onto different and distinguishable points. (b) We randomly sample points in space and project them onto the mesh surface. Left: nearest point projection, projections tend to concentrate on the vertices or edges. Right: proposed dispersed projection, projections are well distributed.

positive for the vertex normal with the face normals of all triangle faces that share this vertex.

Nearest point projection. An obvious way to project a spatial point onto a surface is what we call *nearest point projection*, that is, calculating the point on the surface that is closest to a given point. Animatable NeRF [35] and Neural Actor [24] rely on this way of projection for learning a deformation field and/or utilizing a texture map. While using nearest point projection to obtain \mathbf{s} is straightforward, we note that the mapping $\mathbf{x} \rightarrow \mathcal{X}$ using nearest point projection is not an injection (or one-to-one mapping), thus causing an indistinguishability issue in optimizing NeRF. As the 2D illustration on the left side of Fig. 2(a) shows, the red points on the arc (with the same height h) are all projected onto the blue point, being indistinguishable in the resulting representation. In the case of a 3D mesh, as shown in Fig. 2(b) Left, projected points are concentrated on vertices or edges. The key issue is that NeRF would output the same color and density for points with different spatial positions but the same surface-aligned representation \mathcal{X} , failing to capture sufficient details.

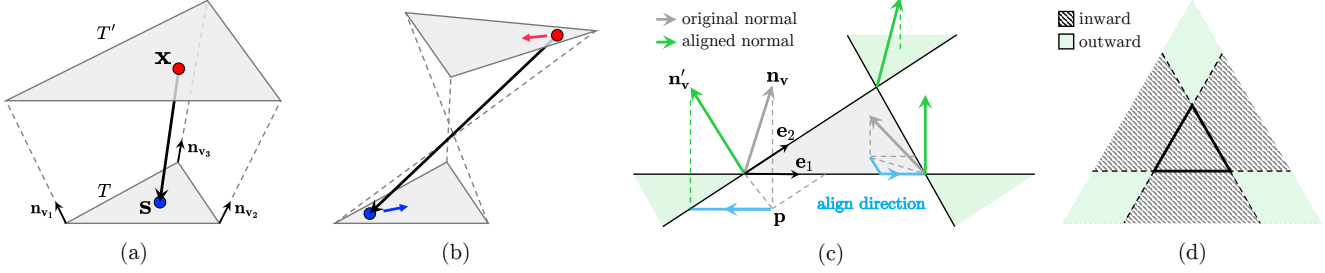


Figure 3. **Illustration of barycentric interpolated projection and vertex normal alignment.** (a) Illustration of barycentric interpolated projection (b) Example of “inverted” projection that occurs when some vertex normals are facing inward. (c) Illustration of vertex normal alignment. (d) Inward and outward regions, viewed from directly above the triangle.

Barycentric interpolated projection. We propose a novel projection method to address the indistinguishability issue in nearest point projection, named *barycentric interpolated projection*. We illustrate the method in Fig. 3(a). For a spatial point \mathbf{x} and a triangle T on the mesh surface, we take the plane that passes through \mathbf{x} and is parallel to T . The intersection of this plane with the three vertex normals forms a parallel triangle T' . Then, we compute the barycentric coordinate of \mathbf{x} in the triangle T' . Finally, using this barycentric coordinate for the triangle T , we can obtain the point \mathbf{s} on the mesh surface through a barycentric interpolation.

Vertex normal alignment. While barycentric interpolated projection works as expected when the vertex normals all appear to be facing *outward*, that is, intuitively, the parallel triangle T' is larger than and therefore able to “cover” T , allowing nearby points to be projected onto T (see Fig. 3(a)). However, when not all the vertex normals face *outward*, as shown in Fig. 3(b), the resulting projection may be “inverted”. To address this, we propose to perform a procedure named *vertex normal alignment*. First, we provide a formal definition of *inward* and *outward*. Let us consider the orthogonal projections of vertex normals of a triangle on a plane. As shown in Fig. 3(c) and (d), we divide the plane into *inward* and *outward* regions. If a vertex normal’s projection falls in the *inward* and *outward* regions, we call the vertex normal *inward* and *outward*, respectively. We align all the *inward* normals of a triangle as follows. First, we consider the orthogonal projection point \mathbf{p} of \mathbf{n}_v . We move \mathbf{n}_v along the directions \mathbf{e}_1 and \mathbf{e}_2 of the two edges until \mathbf{p} reaches the *outward* region, obtaining the aligned direction \mathbf{n}'_v . Finally, we normalize its length to 1 as the aligned vertex normal. Note that vertex normal alignment is performed separately for each triangle, and thus even the same vertex, shared by different triangles, may have differently aligned vertex normals. After vertex normal alignment, the above-mentioned issues of the barycentric interpolated projection can be avoided, and we can obtain more reasonable and distinguishable projection points.

Detailed steps of dispersed projection. We show the detailed steps of projecting a spatial point \mathbf{x} to the mesh surface point \mathbf{s} by combining barycentric interpolated projection and vertex normal alignment.

1. Compute the nearest point $\tilde{\mathbf{s}}$ on the mesh surface.
2. Find a set \mathcal{T} of all triangles containing $\tilde{\mathbf{s}}$. When $\tilde{\mathbf{s}}$ falls inside a triangle, obviously only that triangle contains $\tilde{\mathbf{s}}$; however, when it falls on a vertex or an edge, there will be multiple triangles containing $\tilde{\mathbf{s}}$.
3. Apply vertex normal alignment to all the triangles in \mathcal{T} .
4. For each $T \in \mathcal{T}$, if \mathbf{x} is not inside its parallel triangle T' , remove it from \mathcal{T} . There should be at least one triangle in \mathcal{T} such that \mathbf{x} is inside its T' .
5. Apply barycentric interpolated projection for each $T \in \mathcal{T}$ to obtain the projected surface point $\{\mathbf{s}\}$.
6. Choose the nearest point to \mathbf{x} from $\{\mathbf{s}\}$ as the final projection point.

3.4. Injective mapping to a surface-aligned representation

By projecting \mathbf{x} onto the surface point \mathbf{s} using the proposed dispersed projection, as described in Sec. 3.2, we can compute the surface-aligned representation \mathcal{X} of \mathbf{x} . We emphasize that, using the proposed dispersed projection, the mapping $\mathbf{x} \rightarrow \mathcal{X}$ is an injection under certain conditions. Thus, different spatial points are mapped to different representations, eliminating the indistinguishability issue. The proof can be found in the supplementary materials.

3.5. Surface-aligned neural radiance fields

Our goal is to build neural radiance fields aligned with the mesh surface that can change with the mesh deformation. To this end, we feed the surface-aligned representation \mathcal{X} (Eq. (2)) of a spatial point \mathbf{x} to NeRF as input. Furthermore, we consider the view direction in both the world and

local coordinates. Here, the local coordinates refer to the coordinates formed by the normal, tangent, and bitangent directions on the surface of the posed mesh at s . We denote the view direction in the world coordinates and the local coordinates as \mathbf{d} and \mathbf{d}_l , respectively. We concatenate them together $\mathbf{d}^* = \text{concat}(\mathbf{d}, \mathbf{d}_l)$ and input to the NeRF. This can be written as:

$$F_{\Theta} : (\mathcal{X}, \mathbf{d}^*) \rightarrow (\mathbf{c}, \sigma) \quad (3)$$

where Θ denotes the model parameters. However, surface-aligned representation \mathcal{X} can only capture the motion that strictly follows the body mesh deformation; Eq. (3) cannot model some subtle pose-dependent deformation such as loose clothing. To address this issue, we also condition our model on the skeleton pose parameter p to enable learning the pose-dependent deformation. Specifically, we use an encoder network to encode the pose information as a latent code \mathbf{z}_p and feed it as an additional input to the NeRF:

$$F_{\Theta} : (\mathcal{X}, \mathbf{d}^*, \mathbf{z}_p) \rightarrow (\mathbf{c}, \sigma) \quad (4)$$

This is our proposed surface-aligned NeRF.

3.6. Training

We use volume rendering [16, 28] to synthesize images with a NeRF. Following the previous studies [28, 35, 36], we minimize the per-pixel mean squared error (MSE) between the rendered image and the ground truth image. We use SMPL [26] as our model’s underlying body mesh model with estimated shape and pose parameters. To reduce the impact of an inaccurate SMPL parameter estimation, we optimize the pose parameter of the SMPL simultaneously with the NeRF using Adam optimizer [21]. Additional training details can be found in the supplementary material.

4. Experiments

4.1. Datasets

ZJU-MoCap [36] records human motion using 21 synchronized cameras and uses markerless motion capture to obtain human poses. We use four roughly evenly spaced cameras for training and the remaining cameras for testing novel view synthesis. We divide the motion sequence frames into “training pose” and “unseen pose”, and use the former for training the model and the latter for testing the performance under unseen human pose. We follow [36] for data preprocessing and more training and testing details.

Human3.6M [15] uses four synchronized cameras to record complex human movements and marker-based motion capture to obtain human poses. We use three of the cameras as training view and the remaining one for testing. Similarly, we divide the dataset into the training/unseen pose sequences. For detailed settings, we refer to [35].

Evaluation metrics. Following [28], we use two standard metrics to evaluate the performance of image synthesis: peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM).

4.2. Results on image synthesis

We compare the performance of our approach with that of state-of-the-art methods. For the ZJU-Mocap dataset, we compare with Neural Body [36], which uses structured latent code to model the appearance of the human body. For the Human3.6M dataset, we compare with Animatable NeRF [35], which learns the neural blend weight field and builds a NeRF within a canonical space. In addition, we compare with NARF [30], which takes a similar NeRF approach with an input representation based on the bone coordinates. The results are presented in Tab. 1 and Tab. 2.

Training pose. For the ZJU-MoCap dataset, our approach outperforms [30] and shows a slightly better performance compared to [36], although these studies use per-frame optimization for better modeling of the training pose. For the Human3.6M dataset, our approach outperforms both [30] and [35] by a large margin. The results show that our surface-aligned NeRF can more easily capture the shape and appearance of the human body for the training pose. The qualitative results are presented in Fig. 4. From the synthesized images, we see that the proposed method has better facial details compared to the others.

Unseen pose. For both datasets, the performance of our approach almost consistently outperforms [36], [35], and [30]. Its superior performance may be attributed to its surface-aligned property as it enables fully utilizing the power of SMPL to express body mesh deformations for unseen poses. The qualitative results are presented in Fig. 5. From the synthesized images, we see that the proposed method preserves rich details, such as folds of clothes, even for poses unseen during training, while other methods not only fail to restore details but also produce obvious artifacts.

4.3. Ablation studies

We use sequence “Kick” of the ZJU-MoCap dataset where the actual shape differs significantly from the SMPL mesh estimation (because of the loose hoodie) for an ablation study on the performance of novel view synthesis. The results are shown in Fig. 6 and Tab. 3.

Impact of skeleton pose conditioning. We train a model without the skeleton pose input in Eq. (4), which gives lower PSNR/SSIM results and blurry synthesized images. As discussed in Sec. 3.2, while we assume that the majority of body movements are mesh-following deformations,

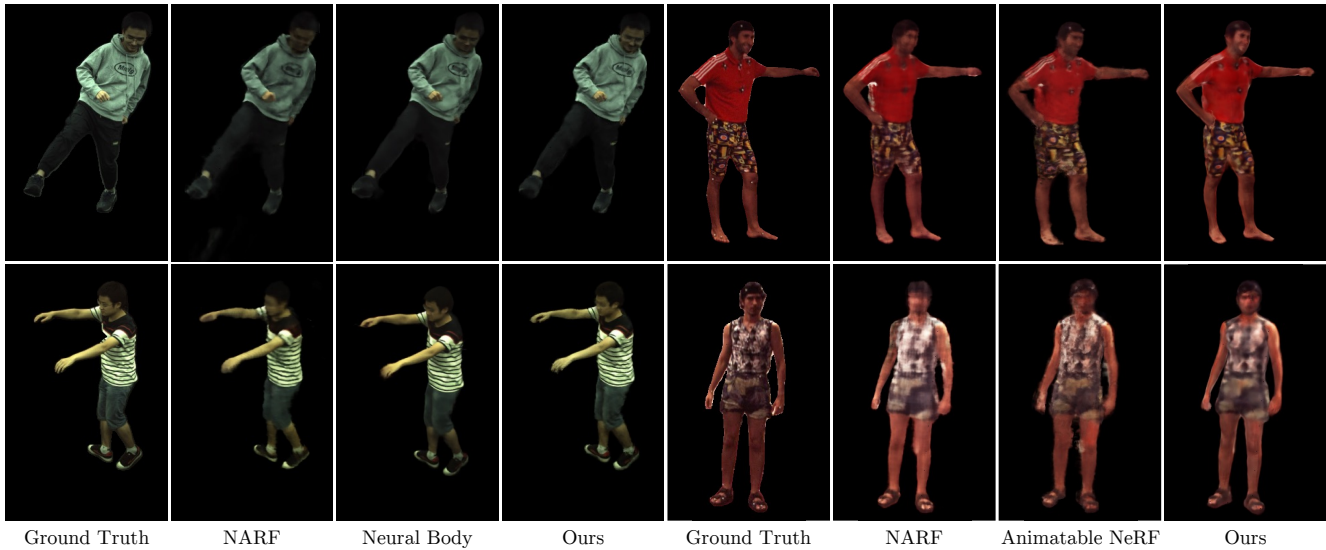


Figure 4. **Qualitative results of novel view synthesis for the training pose.** Left: ZJU-MoCap dataset. Right: Human3.6M dataset.

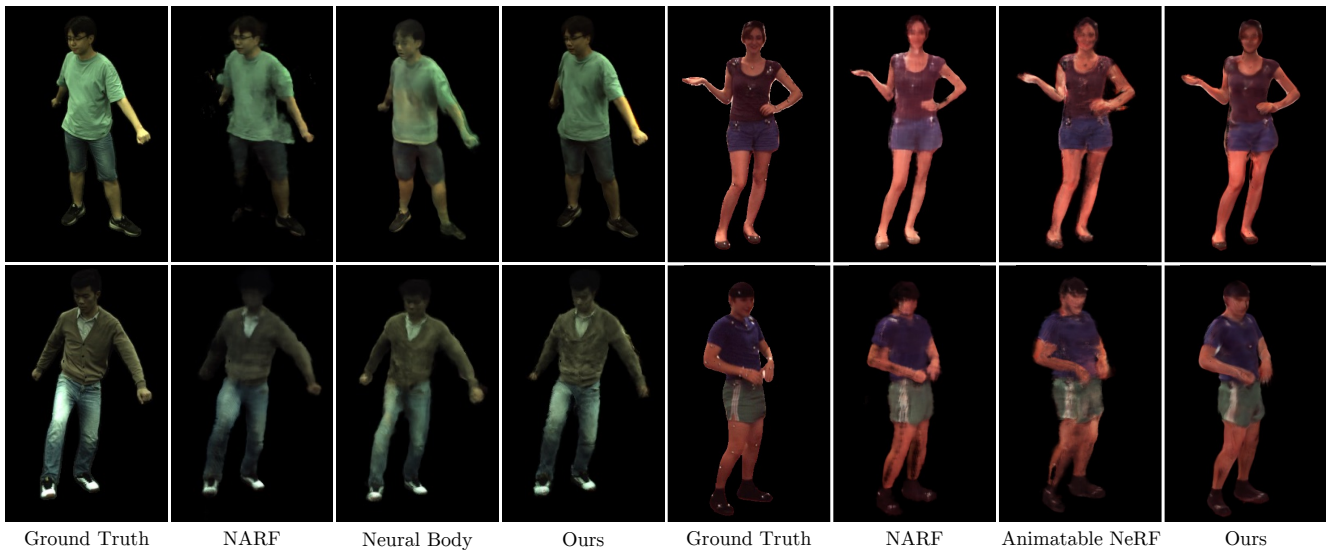


Figure 5. **Qualitative results of novel view synthesis for the unseen pose.** Left: ZJU-MoCap dataset. Right: Human3.6M dataset.

there may also be pose-dependent deformations that cannot be captured by the body mesh. Also, while we do not explicitly model the time-varying deformation components (such as using per-frame embedding in [35,36]), neural networks could implicitly model such components by inferring time from the skeleton pose information.

Impact of the dispersed projection. We use nearest point projection instead of the proposed dispersed projection for training. As shown in Fig. 6, using nearest point projection leads to artifacts and loss of details. As discussed in Sec. 3.3, the dispersed projection addresses the indistinguishability issue arising from nearest point projection.

From the illustration of the nearest point projection (Fig. 2), it is easy to imagine that the farther a point is from the mesh surface, the more likely it is to be projected onto a vertex or an edge by nearest point projection. Based on this observation, we hypothesize that when the estimated mesh is close to the real 3D shape, the indistinguishability may not be so problematic; however, when there are points far from the mesh surface, such as thick clothes or fingers, it could significantly incur the model performance, which may explain the differences.

| | Training pose | | | | | | Unseen pose | | | | | |
|----------|---------------|--------------|--------------|-----------|--------------|--------------|-------------|---------|--------------|-----------|---------|--------------|
| | PSNR | | | SSIM | | | PSNR | | | SSIM | | |
| | NARF [30] | NB [36] | Ours | NARF [30] | NB [36] | Ours | NARF [30] | NB [36] | Ours | NARF [30] | NB [36] | Ours |
| Twirl | 29.38 | 30.56 | 31.32 | 0.967 | 0.971 | 0.974 | 22.20 | 23.95 | 24.33 | 0.872 | 0.905 | 0.908 |
| Taichi | 24.22 | 27.24 | 27.25 | 0.930 | 0.962 | 0.962 | 19.70 | 19.56 | 19.87 | 0.859 | 0.852 | 0.863 |
| Swing1 | 27.53 | 29.44 | 29.29 | 0.929 | 0.946 | 0.946 | 25.43 | 25.76 | 26.27 | 0.912 | 0.909 | 0.927 |
| Swing2 | 27.54 | 28.44 | 28.76 | 0.928 | 0.940 | 0.941 | 24.03 | 23.80 | 24.96 | 0.884 | 0.878 | 0.900 |
| Swing3 | 26.56 | 27.58 | 27.50 | 0.925 | 0.939 | 0.938 | 23.84 | 23.25 | 24.24 | 0.901 | 0.893 | 0.908 |
| Warmup | 25.89 | 27.64 | 27.67 | 0.931 | 0.951 | 0.954 | 24.14 | 23.91 | 25.34 | 0.906 | 0.909 | 0.928 |
| Punch1 | 25.98 | 28.60 | 28.81 | 0.895 | 0.931 | 0.931 | 25.24 | 25.68 | 27.30 | 0.877 | 0.881 | 0.905 |
| Punch2 | 24.78 | 25.79 | 26.08 | 0.915 | 0.928 | 0.929 | 22.58 | 21.60 | 23.08 | 0.885 | 0.870 | 0.890 |
| Kick | 26.42 | 27.59 | 27.77 | 0.913 | 0.926 | 0.927 | 23.53 | 23.90 | 24.43 | 0.872 | 0.870 | 0.889 |
| average. | 26.48 | 28.10 | 28.27 | 0.926 | 0.944 | 0.945 | 23.41 | 23.49 | 24.42 | 0.885 | 0.885 | 0.902 |

Table 1. Results of the ZJU-MoCap dataset [36] in terms of PSNR and SSIM. Higher is better. “NB” means Neural Body.

| | Training pose | | | | | | Unseen pose | | | | | |
|----------|---------------|---------|--------------|--------------|---------|--------------|--------------|---------|--------------|--------------|---------|--------------|
| | PSNR | | | SSIM | | | PSNR | | | SSIM | | |
| | NARF [30] | AN [35] | Ours | NARF [30] | AN [35] | Ours | NARF [30] | AN [35] | Ours | NARF [30] | AN [35] | Ours |
| S1 | 21.41 | 22.05 | 23.71 | 0.891 | 0.888 | 0.915 | 20.19 | 21.37 | 22.67 | 0.864 | 0.868 | 0.890 |
| S5 | 25.24 | 23.27 | 24.78 | 0.914 | 0.892 | 0.909 | 23.91 | 22.29 | 23.27 | 0.891 | 0.875 | 0.881 |
| S6 | 21.47 | 21.13 | 23.22 | 0.871 | 0.854 | 0.881 | 22.47 | 22.59 | 23.23 | 0.883 | 0.884 | 0.888 |
| S7 | 21.36 | 22.50 | 22.59 | 0.899 | 0.890 | 0.905 | 20.66 | 22.22 | 22.51 | 0.876 | 0.878 | 0.898 |
| S8 | 22.03 | 22.75 | 24.55 | 0.904 | 0.898 | 0.922 | 21.09 | 21.78 | 23.06 | 0.887 | 0.882 | 0.904 |
| S9 | 25.11 | 24.72 | 25.31 | 0.906 | 0.908 | 0.913 | 23.61 | 23.72 | 23.84 | 0.881 | 0.886 | 0.889 |
| S11 | 24.35 | 24.55 | 25.83 | 0.902 | 0.902 | 0.917 | 23.95 | 23.91 | 24.19 | 0.885 | 0.889 | 0.891 |
| average. | 23.00 | 23.00 | 24.28 | 0.898 | 0.890 | 0.909 | 22.27 | 22.55 | 23.25 | 0.881 | 0.880 | 0.892 |

Table 2. Results of the Human3.6M dataset [15] in terms of PSNR and SSIM. Higher is better. “AN” means Animatable NeRF.

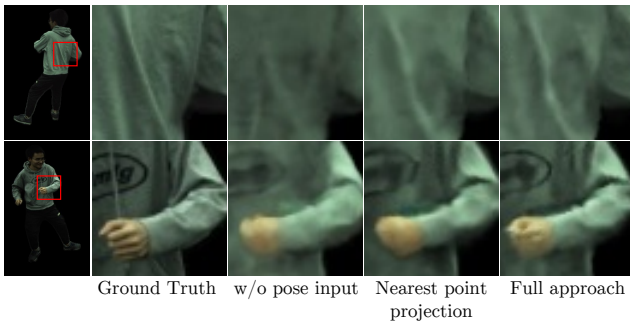


Figure 6. Qualitative results of ablation studies using sequence “Kick” of ZJU-MoCap dataset. “w/o pose input” indicates that skeleton pose is not used as input to NeRF. “Nearest point projection” means we use nearest point projection instead of proposed dispersed projection to calculate the surface-aligned representation.

| | PSNR | SSIM |
|--------------------------|--------------|--------------|
| w/o pose input | 26.58 | 0.915 |
| Nearest point projection | 27.38 | 0.923 |
| Full approach | 27.77 | 0.927 |

Table 3. Results of ablation studies using sequence “Kick” of ZJU-MoCap dataset in terms of PSNR and SSIM. Higher is better. Also refer to Fig. 6.

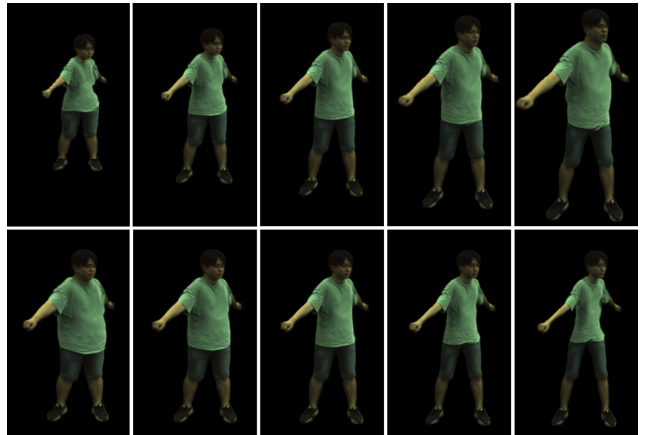


Figure 7. Qualitative results of body shape control. Top: change PC1 from -4 to $+4$. Bottom: change PC2 from -4 to $+4$.

4.4. Body shape control

Because the proposed NeRF is aligned to the body mesh surface, we can control the body shape of modeled humans by manipulating the body mesh surface. Specifically, the SMPL has 10 shape parameters obtained by principal component analysis (PCA), and by controlling them, we can obtain meshes of different body shapes. Fig. 7 shows the synthesized images after we changed the first and second principal components, PC1 and PC2.

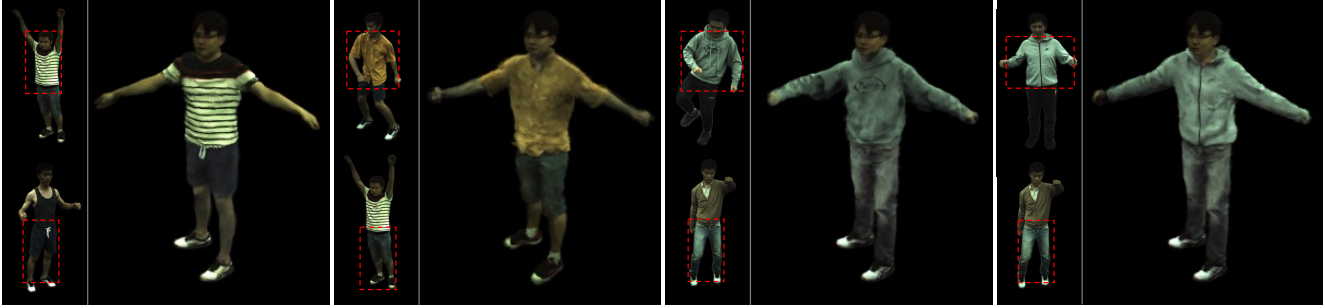


Figure 8. **Qualitative results of changing clothes.** Since the surface-aligned property of our method, we can use different trained NeRFs for different mesh surface areas. The small images on the left indicate the source appearance and the large image on the right indicates the synthesized image with novel appearance combination.

4.5. Changing clothes

As we can replace a part of the mesh texture with another texture, our method can also replace NeRF according to the different surface areas to achieve a similar effect of “changing clothes”.

Suppose that we already have several proposed surface-aligned NeRF models trained on different subjects. We first segment the SMPL mesh according to the body parts, such that each triangle face belongs to a category (*e.g.*, head, upper body, and lower body). We then decide which subject’s appearance to use for each category (*e.g.*, we use the upper body of subject 1 and the lower body of subject 2). For a query point \mathbf{x} , we project it onto the mesh surface point \mathbf{s} using the proposed dispersed projection and find the category to which the \mathbf{s} belongs. Given this category, we compute the color and the density of point \mathbf{x} using the trained NeRF model of the corresponding subject. Finally, we can render the novel synthesized image with a combination of the appearance of multiple subjects. The synthesized results are shown in Fig. 8.

5. Discussion

5.1. Limitations

Currently, our method relies on a relatively accurate mesh estimation. Although we can alleviate inaccuracies by optimizing the SMPL parameters during training, some details that cannot be accurately represented through SMPL, such as hand pose, may lead to artifacts because points cannot be projected correctly onto the mesh surface. A possible solution is to replace SMPL with parametric human models in more detail, such as with SMPL-X [34].

The proposed surface-aligned representation using only one surface point may have potential issues that, in some specific poses such as when the arm and the body are extremely close, some points that are projected to the body in the training pose for representing the body information, may be projected to the arm in the novel pose setting, caus-

ing that point cannot correctly represent the body information. It would be interesting to introduce more information, such as interrelationships of spatial points with each body part as similar to [30] for further improvement.

Our dispersed projection assumes a watertight mesh with vertex normals all forming acute angles with neighboring face normals, which may be restrictive when applied to more complicated meshes. Designing a more flexible projection method without the indistinguishability issue would broaden the applicability of surface-aligned representation.

5.2. Future work

The core idea of our method is to build a NeRF on the estimated mesh surface. Therefore, our method can be generalized to objects with corresponding 3D mesh, not limited to the human body, with the help of the existing 3D mesh reconstruction methods from images [7, 17–19, 25] or videos [11, 22, 31, 38, 43]. Furthermore, combined with interactive 3D mesh deformation methods [12, 14, 20], our method can enable interactive manipulations of NeRF. We expect the idea of combining the controllability of mesh and photorealistic rendering of NeRF can be used for more visual applications in the future.

6. Conclusion

In this paper, we present novel surface-aligned neural radiance fields for a controllable 3D human synthesis. The proposed dispersed projection method transforms spatial points into distinguishable surface points and signed heights, which shows high compatibility with the proposed surface-aligned NeRF. Our method not only shows a better generalization performance in the novel human pose situations, but it also supports explicit controls such as body shape change and clothes change.

Acknowledgements: We thank Hiroharu Kato for helpful discussions and comments.

References

- [1] 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, 2
- [2] 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] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In *ICCV*, 2019. 1, 2
- [4] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. Scape: Shape completion and animation of people. *ACM Trans. Graph.*, 2005. 1, 2
- [5] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *ECCV*, 2020. 2
- [6] Joel Carranza, Christian Theobalt, Marcus A. Magnor, and Hans-Peter Seidel. Free-viewpoint video of human actors. *ACM Trans. Graph.*, 2003. 1
- [7] Wenzheng Chen, Jun Gao, Huan Ling, Edward Smith, Jaakko Lehtinen, Alec Jacobson, and Sanja Fidler. Learning to predict 3d objects with an interpolation-based differentiable renderer. In *NeurIPS*, 2019. 8
- [8] Boyang Deng, John P. Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey E. Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. NASA neural articulated shape approximation. In *ECCV*, 2020. 2
- [9] Yilun Du, Yanan Zhang, Hong-Xing Yu, Joshua B. Tenenbaum, and Jiajun Wu. Neural radiance flow for 4d view synthesis and video processing. In *ICCV*, 2021. 2
- [10] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *ECCV*, 2018. 2
- [11] Philipp Henzler, Jeremy Reizenstein, Patrick Labatut, Roman Shapovalov, Tobias Ritschel, Andrea Vedaldi, and David Novotny. Unsupervised learning of 3d object categories from videos in the wild. In *CVPR*, 2021. 8
- [12] Jing Hua and Hong Qin. Free-form deformations via sketching and manipulating scalar fields. In *ACM Symposium on Solid Modeling and Applications*, 2003. 8
- [13] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. ARCH: Animatable Reconstruction of Clothed Humans. In *CVPR*, 2020. 2
- [14] Takeo Igarashi, Satoshi Matsuoka, and Hidehiko Tanaka. Teddy: A sketching interface for 3d freeform design. In *ACM SIGGRAPH*, 1999. 8
- [15] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2014. 5, 7
- [16] James T. Kajiya. The rendering equation. *SIGGRAPH Comput. Graph.*, 1986. 5
- [17] Angjoo Kanazawa, Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *ECCV*, 2018. 8
- [18] Hiroharu Kato and Tatsuya Harada. Learning view priors for single-view 3d reconstruction. In *CVPR*, 2019. 8
- [19] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *CVPR*, 2018. 8
- [20] Youngihm Kho and Michael Garland. Sketching mesh deformations. In *ACM SIGGRAPH*, 2005. 8
- [21] Diederick P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5
- [22] Xueting Li, Sifei Liu, Shalini De Mello, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, and Jan Kautz. Online adaptation for consistent mesh reconstruction in the wild. In *NeurIPS*, 2020. 8
- [23] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *CVPR*, 2021. 2
- [24] 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 Trans. Graph.(ACM SIGGRAPH Asia)*, 2021. 1, 2, 3
- [25] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. *ICCV*, 2019. 8
- [26] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 2015. 1, 2, 5
- [27] 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. 1, 2
- [28] 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. 1, 2, 5
- [29] Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. SiCloPe: Silhouette-Based Clothed People. In *CVPR*, Long Beach, CA, 2019. 2
- [30] Atsuhiko Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *ICCV*, 2021. 1, 2, 5, 7, 8
- [31] David Novotny, Diane Larlus, and Andrea Vedaldi. Learning 3d object categories by looking around them. In *ICCV*, 2017. 8
- [32] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. *ICCV*, 2021. 2
- [33] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.(ACM SIGGRAPH Asia)*, 2021. 2

- [34] 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. 1, 8
- [35] 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, 5, 6, 7
- [36] 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. 1, 2, 5, 6, 7
- [37] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural Radiance Fields for Dynamic Scenes. In *CVPR*, 2021. 2
- [38] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *ICCV*, 2021. 8
- [39] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. *ICCV*, 2019. 2
- [40] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *CVPR*, June 2020. 2
- [41] 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 dynamic scene from monocular video. In *ICCV*, 2021. 2
- [42] Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović. Articulated mesh animation from multi-view silhouettes. *ACM Trans. Graph.*, 2008. 1
- [43] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Huiwen Chang, Deva Ramanan, William T Freeman, and Ce Liu. Lasr: Learning articulated shape reconstruction from a monocular video. In *CVPR*, 2021. 8
- [44] Ze Yang, Shenlong Wang, Sivabalan Manivasagam, Zeng Huang, Wei-Chiu Ma, Xinchen Yan, Ersin Yumer, and Raquel Urtasun. S3: Neural shape, skeleton, and skinning fields for 3d human modeling. In *CVPR*, 2021. 2