H-NeRF: Neural Radiance Fields for Rendering and Temporal Reconstruction of Humans in Motion

Hongyi Xu
Google Research
hongyixu@google.com

Thiemo Alldieck
Google Research
alldieck@google.com

Cristian Sminchisescu
Google Research
sminchisescu@google.com

Abstract

We present H-NeRF, neural radiance fields for rendering and temporal (4D) reconstruction of a human in motion as captured by a sparse set of cameras or even from a monocular video. Our NeRF-inspired approach combines ideas from neural scene representation, novel-view synthesis, and implicit statistical geometric human representations. H-NeRF allows to accurately synthesize images of the observed subject under novel camera views and human poses. Instead of learning a radiance field in empty space, we attach it to a structured implicit human body model, represented using signed distance functions. This allows us to robustly fuse information from sparse views and, at test time, to extrapolate beyond the observed poses or views. Moreover, we apply geometric constraints to co-learn the structure of the observed subject (including both body and clothing) and to regularize the radiance field to geometrical plausible solutions. Extensive experiments on multiple datasets demonstrate the robustness and accuracy of our approach and its generalization capabilities beyond the sparse training set of poses and views.

1 Introduction

Enabling free-viewpoint video of a human subject in motion is a challenging problem with many applications. Our work is motivated by the breadth of transformative 3D applications that would become possible, including immersive visualization of photos, virtual clothing and try-on modeling, fitness, or AR and VR for improved communication or collaboration. So far static scenes have been the primary subject of research. In pursuing realistic novel-view synthesis two schools of thought have been established: 1) 3D reconstruction methods aim to recover the geometry of the observed scene as accurately as possible before being displayed from novel views using classical rendering pipelines [31, 43]. 2) Image-based rendering techniques [8, 11, 41] and very recently neural radiance fields [30] primarily aim for image production quality without explicitly aiming at an accurate 3D geometric model. While these techniques sometimes explicitly or implicitly reconstruct the scene geometry as well, the geometry is not guaranteed to accurately resemble the true scene geometry. We argue that novel-view rendering and reconstruction are two sides of the same coin, and reliable viewpoint generalization, especially given relatively few input views, would require good quality for both. To this end, we propose a unified model in order to support robust both reconstruction and photo-realistic rendering. Dynamic scenes, especially those capturing a human in motion, add a new level of complexity to the problem: while static scenes can be observed from many views by a camera moving through the scene, any configuration of a dynamic scene is typically observed only from sparse views. Moreover, the scene geometry and its appearance may change drastically over time. To cope with the fewer views, some methods integrate scene knowledge over time by warping observations into a common reference frame [36, 38]. At test time, the information is warped back to the desired state and rendered from a novel view. Extrapolating to unseen motion, however, remains challenging. For scenes capturing people, this means that only poses seen during training can be rendered at test time. For some applications, however, rendering the subject under a large range of motions or in novel poses is desired. To make generalisation over poses and views possible, we rely on additional problem domain knowledge in the form of the human body model imGHUM [5]. imGHUM is an implicit signed distance function (SDF) conditioned on generative shape and pose

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

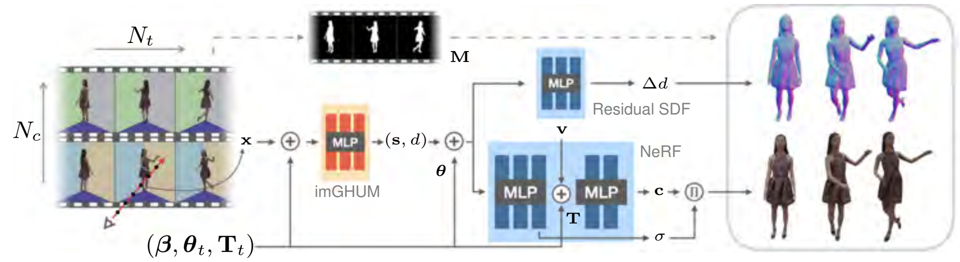

Figure 1: Overview of H-NeRF. Given a collection of images of a human performer collected from a sparse set of calibrated cameras, we train a geometry-aware neural radiance field by co-learning a deformable signed distance field. First, we estimate the body shape $\boldsymbol{\beta}$ and track the articulated pose $\boldsymbol{\theta}, \mathbf{T}$ using the implicit statistical human body model imGHUM (orange). imGHUM encodes all 3D spatial points $\mathbf{x}$ across frames with a 4D point descriptor $(\mathbf{s}, d)$ referencing a canonical frame. Using the foreground mask $\mathbf{M}$, we co-train a residual SDF and a NeRF network in that canonical space which integrate all visible 2D observations into a consistent implicit 3D geometry Δd and its view-depended $(\mathbf{v})$ radiance representation $(\mathbf{c}, \sigma)$. The trainable residual SDF and NeRF networks (in blue) are conditioned on the body pose $\boldsymbol{\theta}$ and the root transformation $\mathbf{T}$ to model pose-dependent geometric and appearance variations. Our framework supports both accurate 3D geometric reconstruction and free-viewpoint rendering, and generalizes well to novel views, shapes, and poses.

codes learned from a large corpus of dynamic human scans. In this work, imGHUM is used as the common reference frame for a neural radiance field and as a structured prior for robust reconstruction. Additionally, since imGHUM can represent a broad distribution of statistically valid human poses and shapes, we can render the reconstructed subject in novel poses and even with modified body shapes. In summary, our system supports photo-realistic free-view point rendering of a human subject as observed from sparse cameras. By conditioning on an implicit human body model, we can render novel body poses and shapes. Our carefully chosen losses ensure not only image quality but also plausible temporal reconstructions that can be used for the free-viewpoint visualisation of human performance capture.

2 Related Work

Human performance capture is the process of reconstructing the dynamic (4D) geometry of a human subject as observed in one or multiple synchronized views. Pioneering methods deform a rigged template mesh according to silhouettes observed in multiple views [7, 51], estimated skeleton key-points [13], or image correspondences [10]. Later systems relying on RGB-D video streams [6] or monocular capture [4, 3, 18, 53, 17, 16] have also been presented. All these methods rely either on a pre-scanned template mesh or on a body model that is deformed to fit the image evidence.

Neural representation for view synthesis. With the advent of neural networks researchers have begun to explore alternative solutions to represent a scene to support novel view synthesis and photo-realistic rendering; we refer the reader to [49] for a survey. Various forms of scene representation have been explored. Some methods use voxel grids to embed the scene [24, 46]. For rendering, the voxel grid is probed by shooting a ray and by linearly interpolating voxel values. These samples are then transformed into color values using a neural network. Other researchers have proposed neural textures [50, 44] that can be rendered with view-dependent effects or texture synthesis networks [23] for realistic rendering of meshes. In a similar spirit, other methods render point clouds, where each point carries local appearance information [27, 2]. Riegler and Koltun [41] reproject features from nearby views and rely on a SfM reconstruction scaffold for novel view synthesis. With the recent advent of implicit function networks [26, 35, 9, 28], such 3D representations have been explored for rendering and view synthesis with great success. In contrast to discrete voxel representations, textures, or point clouds, these methods represent the scene as a continuous function and thus are not bound to a specific image or volume resolution. In the pioneering work by Sitzmann et al. an implicit function produces features rendered using a neural renderer [47]. Follow up methods focus on 3D geometric reconstruction [45], and use 2D supervision [32, 54].

Neural Radiance Fields (NeRFs) are a recent approach to represent scenes for novel view synthesis. Mildenhall et al. [30] introduce Neural Radiance Fields as fully-connected neural networks representing a scene, whose input is a spatial query point and a viewing direction. The output is a volume density and the emitted radiance at the query location in the direction of the viewer. By ray-tracing using this simple representation, one can generate photo-realistic images from novel views. Despite the excellent quality of results one drawback is the slow rendering time. To this end, researchers have presented faster versions that e.g. transform the radiance field into more efficient sparse grids [19] or remove the dependence on the view-direction during rendering by estimating a spherical harmonic representation of the radiance function [55]. Other derivatives improve fidelity or rendering time by tackling ambiguities in the original formulation [58], spatial decomposition into multiple NeRFs [40], or combining NeRF with sparse voxel fields [22]. Initial work adapting NeRF for dynamic scenes has been presented as well. Park et al. [36] produce "Nerfies" (NeRF-Selfies) from videos where subjects carefully move a camera around their head. The scene information is fused by warping query points into a canonical reference frame. Similarly, Pumarola et al. [38] produce dynamic NeRFs from synthetic animation data. Related to our approach, some methods integrate human body models to fuse information over time. A-NeRF [48] uses a skeleton to rigidly transform NeRF features to refine estimated 3D poses. A similar approach is followed in NARF [33] for view synthesis. Most related, Neural Body [37] attaches learnable features to the vertices of a SMPL body model [25]. These features are processed with a sparse 3D convolutional network, where the output forms a neural radiance field. In contrast to our approach, the resolution is bounded by the spatial resolution of the 3D conv net and no geometric supervision is performed. We highlight further differences in §5.

3 Background

Given a collection of images capturing a dynamic scene of a human in motion, seen from calibrated sparse camera views (in the limit a single monocular camera, as we will show), we aim to learn both the detailed temporal geometry of the human in motion, and enable rendering of the sequence from novel camera views and for novel human poses. To this end, our work unifies two main methodologies: 1) implicit 3D human representations and 2) volumetric radiance fields. In this section, we provide the relevant background on both representations, which we co-learn in a unified framework.

Neural Radiance Fields. A neural radiance field (NeRF) [30] represents a 3D scene as a continuous function of color volume densities. More specifically, the model consists of a neural network function F_ω that maps a 3D spatial point $\mathbf{x} \in \mathbf{R}^3$ and a viewing direction $\mathbf{v} \in \mathbf{R}^3$ to a volume density $\sigma \in \mathbf{R}^+$ and a radiance $\mathbf{c}(\mathbf{x}, \mathbf{v}) \in \mathbf{R}^3$ emitted towards the viewer. In practice, NeRF encodes the inputs $\mathbf{x}$ and $\mathbf{v}$ using a sinusoidal positional encoding $\boldsymbol{\gamma} : \mathbf{R}^3 \to \mathbf{R}^{3+6m}$ that projects a coordinate vector into a high-dimensional space using a set of sine and cosine functions of m increasing frequencies. Given a ray $\mathbf{r} = \mathbf{o} + s\mathbf{v}$ with N samples $\{\mathbf{x}\}$ originating from a camera location $\mathbf{o}$, NeRF integrates radiance values along the ray by means of alpha blending. The pixel/ray color is approximated with numerical quadrature [34]:

$$\mathbf{C}(\mathbf{r}) = \sum_{i=1}^{N} \alpha(\mathbf{x}_i) \prod_{j<i} (1 - \alpha(\mathbf{x}_j)) \mathbf{c}(\mathbf{x}_i, \mathbf{v}), \tag{1}$$

$$\alpha(\mathbf{x}_i) = 1 - \exp\big(-\sigma(\mathbf{x}_i)\delta_i\big), \tag{2}$$

where $\alpha(\mathbf{x}_i)$ is the transparency by accumulating transmittance along the ray, and $\delta_i = |\mathbf{x}_{i+1} - \mathbf{x}_i|$ is the distance between adjacent samples. The NeRF function F_w is fully differentiable and its network parameters w can be optimized using an image reconstruction loss [30]. An approximate 3D scene geometry $\mathbf{S}_F = \{\mathbf{x} | \sigma(\mathbf{x}) = \sigma_h\}$ can be extracted from the trained opacity field via Marching Cubes [21] at a density threshold σ_h.

Implicit Generative Human Models. Implicit human body surfaces are typically represented as the decision boundary of either binary occupancy classifiers [12, 42, 29] or signed distance functions [15, 5]. More specifically, our work builds upon the SOTA statistical implicit human model imGHUM [5] $H_\omega : (\mathbf{T}^{-1}\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\theta}) \to (d, \mathbf{s})$ that maps a 3D spatial point $\mathbf{x}$, unposed with root joint transformation $\mathbf{T} \in \mathbf{R}^{4\times3}$, to its signed distance value $d \in \mathbf{R}$ with respect to a body surface parameterized with body shape $\boldsymbol{\beta} \in \mathbf{R}^{16}$ and articulated pose $\boldsymbol{\theta} \in \mathbf{R}^{118}$. In addition to d, imGHUM returns implicit continuous semantics $\mathbf{s} \in \mathbf{R}^3$ of the query point which correspond the 3D coordinate of the nearest surface point defined on a canonical surface. We refer to the original paper [5] for details. Essentially, imGHUM builds a human-centric 4D semantic descriptor $(d, \mathbf{s})$ for all spatial points around the parameterized human body. imGHUM is trained from a large collection of human scans with diverse body shapes and poses, sharing the same generative shape and pose latent code with the mesh-based statistical human model GHUM [52]. The 3D implicit articulated human body $\mathbf{S}_H(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{T}) = \{\mathbf{x} | d(\mathbf{x}) = 0\}$ is defined by the the zero-isosurface of the signed distance field.

4 Method

In this section, we describe our main contribution H-NeRF, a novel neural network (fig. 1) that exploits both the power of volumetric radiance fields for learning complex scene structure and appearance, as well as implicit signed distance functions for accurate geometric reconstruction. Further, H-NeRF utilizes imGHUM, an implicit model of articulated human pose and shape, as a rich geometric prior, to integrate scene information over time, and to parameterize human articulation. Co-learning both a radiance field and a signed distance function of scene geometry consistently in a unified framework, enables 1) accurate 3D geometric reconstruction and volume rendering of a dynamic human on motion; 2) low demand on the number of training camera views; 3) consistent integration of 2D observations over time; and 4) good generalization capability to novel camera views, human shapes, and poses.

Given a sparse calibrated multi-view video of a human in motion, our task is to generate free-viewpoint video of the observed person and to reconstruct the underlying 4D geometry. We denote the input image collection with a resolution of $w \times h$ pixels as $\{\mathbf{I}_t^c \in \mathbf{R}^{w \times h \times 3} | c = 1, \ldots, N_c, t = 1, \ldots, N_t\}$, where c is the camera index, N_c is the number of cameras, t is the frame index, and N_t is the number of frames. For each image, we apply [14] to obtain the binary foreground human mask $\mathbf{M}_t^c \in \mathbf{R}^{w \times h}$. In addition, we obtain a temporal consistent imGHUM shape latent code $\boldsymbol{\beta}$ and pose codes $(\boldsymbol{\theta}_t, \mathbf{T}_t)$ at each frame index t by optimizing multi-view 2D keypoint and body segmentation losses [57]. We refer to our supplemental material for the detailed imGHUM fitting process.

In the sequel we first explain our adaptations to NeRF for the static case. imGHUM serves as a scene prior that helps to structure the observed scene and to learn more accurate geometry. We continue by explaining which additional changes are needed in the dynamic case. Hereby, imGHUM provides spatial correspondences across frames and is used as the common reference frame where we fuse information from different views and time instances.

4.1 Static Semantic Human NeRF

We first formulate H-NeRF for a human capture at a single moment in time ($N_t = 1$). From multi-view 2D observations, we co-learn a radiance field $F_\omega : \mathbf{x}, \mathbf{v} \to (\mathbf{c}, \sigma)$ for free-viewpoint rendering, and a signed distance function $\hat{H}_\omega : \mathbf{x} \to \hat{d}$ for 3D geometric reconstruction. We use $\hat{H}_\omega$ for the dressed subject to distinguish it from the minimally-dressed imGHUM body SDF H_ω.

Like NeRF, we formulate an L_1 image reconstruction loss to optimize F_w as

$$\mathcal{L}_{\text{rec}} = \sum_{\mathbf{r} \in \mathcal{R}} \|\bar{\mathbf{C}}(\mathbf{r}) - \mathbf{C}(\mathbf{r})\|_1, \tag{3}$$

where $\mathcal{R}$ denotes the batched set of all pixels/rays, and $\bar{\mathbf{C}}(\mathbf{r}), \mathbf{C}(\mathbf{r})$ are observed and rendered pixel color (cf. (1)), respectively. However, under sparse training camera views, the volumetric radiance field is not well regularized, leading to poor generalization to novel camera views, cf. fig. 2. Specifically, we observe that the model has difficulty in correctly representing the scene and separating the person from background. The model fails to learn a semantically meaningful opacity field, i.e. $\alpha = 0$ in the free space and 1 in the occupied space.

Coarse Scene Structuring. Using imGHUM fits for all training images, we can spatially locate the person in 3D space. We define a 3D bounding box $\mathbf{B} \in [\underline{\mathbf{S}_H} - \epsilon, \overline{\mathbf{S}_H} + \epsilon]$ around the detected person, where $\underline{\mathbf{S}_H}, \overline{\mathbf{S}_H}$ are the minimal and maximal coordinates of the human body surface $\mathbf{S}$ and ϵ is a spatial margin reserved for geometry not modeled by imGHUM. All radiance points that render the person should now reside inside the bounding box, leading to our 3D segmentation loss:

$$\tilde{\mathbf{C}}(\mathbf{r}) = \sum_{i=1}^{N} b(\mathbf{x}_i)\alpha(\mathbf{x}_i) \prod_{j<i} (1 - b(\mathbf{x}_j)\alpha(\mathbf{x}_j))\mathbf{c}(\mathbf{x}_i, \mathbf{v}), \tag{4}$$

$$\mathcal{L}_{\text{mask}} = \sum_{\mathbf{r} \in \mathcal{R}} \left\| \mathbf{M}(\mathbf{r}) \left(\bar{\mathbf{C}}(\mathbf{r}) - \tilde{\mathbf{C}}(\mathbf{r}) \right) \right\|_1, \tag{5}$$

where $b(\mathbf{x}_i)$ is 0 if outside of $\mathbf{B}$ and 1 otherwise, and $\mathbf{M}(\mathbf{r})$ is the image mask. We apply the mask here so the loss is only applied to structure that belongs to the subject and not to other scene geometry.

Unifying SDF with NeRF. After structuring the scene coarsely, we now couple the radiance field with implicit signed distance-based 3D reconstruction to further regularize the opacity distribution.

A signed distance function describing the detected person in the image naturally comes with a 3D classifier for all spatial points, where $\hat{d}(\mathbf{x}_i) > 0$ means $\mathbf{x}_i$ is in free space whereas $\mathbf{x}_i$ lies within the subject when $\hat{d}(\mathbf{x}_i) <= 0$. We rely on this observation to co-learn a SDF of the performer as an inductive bias for the radiance field. To this end, we introduce a pseudo alpha value $\dot{\alpha}(\mathbf{x}_i) = \phi(\gamma \hat{d}(\mathbf{x}_i))$ where ϕ is a Sigmoid activation function and γ controls the sharpness of the boundary. To refine the NeRF opacity semantics, especially for the volume around the human $\mathbf{B}$, we formulate two losses:

$$\hat{\mathbf{C}}(\mathbf{r}) = \sum_{i=1}^{N} \hat{\alpha}(\mathbf{x}_i) \prod_{j<i} (1 - \hat{\alpha}(\mathbf{x}_j)) \mathbf{c}(\mathbf{x}_i, \mathbf{v}), \quad \hat{\alpha}(\mathbf{x}_i) = b(\mathbf{x}_i)\dot{\alpha}(\mathbf{x}_i) + (1 - b(\mathbf{x}_i))\alpha(\mathbf{x}_i), \tag{6}$$

$$\mathcal{L}_{\text{blend}} = \sum_{\mathbf{r} \in \mathcal{R}} \big(\|\mathbf{M}(\mathbf{r})\big(\bar{\mathbf{C}}(\mathbf{r}) - \hat{\mathbf{C}}(\mathbf{r})\big)\|_1 + \eta \|(1 - \mathbf{M}(\mathbf{r}))\big(\bar{\mathbf{C}}(\mathbf{r}) - \hat{\mathbf{C}}(\mathbf{r})\big)\|_1 \big), \tag{7}$$

$$\mathcal{L}_{\text{geom}}(\mathbf{r}) = \sum_{i=1}^{N} b(\mathbf{x}_i) \text{BCE}\big(\phi\big(\lambda(\sigma_h - \sigma(\mathbf{x}_i))\big), \dot{\alpha}(\mathbf{x}_i)\big), \tag{8}$$

where $\hat{\mathbf{C}}(\mathbf{r})$ is the rendered pixel color with blended alpha values $\hat{\alpha}$ replacing NeRF alpha values α with SDF-based pseudo alpha $\dot{\alpha}$ for all ray points inside the bounding box $\mathbf{B}$. The first term in $\mathcal{L}_{\text{blend}}$ requires the rendered pixel color for all the intersecting rays within the human mask to come from the surface, whereas the second term assumes that all background color is formed from ray samples outside of $\mathbf{B}$. We set $\eta = 1$ when no other geometry than the person is inside $\mathbf{B}$ and tune η down if the assumption is violated (e.g. a person standing on a floor). The term $\mathcal{L}_{\text{geom}}$ uses the binary cross entropy loss to couple the NeRF surface boundary with the zero-isosurface of the signed distance function describing the subject. While the coupling terms given by (7) and (8) act as strong priors for the opacity distribution, during test time, we still rely on volumetric radiance rendering with learned NeRF alpha values α to support transparency effects and complex geometry such as hair.

Image-based SDF learning. Learning the signed distance field $\hat{H}_\omega$ from scratch using a sparse set of training images still remains challenging. The model often fails to reconstruct reasonable human geometry, even when using our coupling losses (7) and (8). We therefore leverage imGHUM as an inner layer for the target person and combine it with a light-weight residual SDF network $\Delta H_\omega : \mathbf{x} \to \Delta d$. The residual SDF models all surface details, including hair and clothing, that are not represented by imGHUM. The final signed distance for $\mathbf{x}_i$ becomes $\hat{d}(\mathbf{x}_i) = d(\mathbf{x}_i|\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{T}) + \Delta d(\mathbf{x}_i)$. Given the training images, we learn the personalized residual SDF with our coupling losses, given by (7) and (8), and additionally apply geometric regularization:

$$\mathcal{L}_{\text{seg}} = \sum_{\mathbf{r} \in \mathcal{R}} \text{BCE}(\mathbf{M}(\mathbf{r}), \hat{d}_{\text{min}}(\mathbf{r})), \tag{9}$$

$$\mathcal{L}_{\text{eik}}(\mathbf{r}) = \sum_{i=1}^{N} b(\mathbf{x}_i)(\|\nabla_{\mathbf{x}_i} \hat{d}(\mathbf{x}_i)\|_2 - 1)^2, \quad \mathcal{L}_{\text{reg}} = \|\psi(\Delta d)\|_1, \tag{10}$$

where $\hat{d}_{\text{min}}(\mathbf{r})$ denotes the minimal signed distance for all sampled ray points and ψ is the ReLU activation function. The term $\mathcal{L}_{\text{seg}}$ ensures that if a pixel is inside the human segmentation mask, there should be at least one intersection between the ray and the 3D human surface and therefore $\hat{d}_{\text{min}}(\mathbf{r})$ should be non-positive. Otherwise, all ray samples should have positive signed distances. With $\mathcal{L}_{\text{eik}}$, we enforce the composite SDF $\hat{d}$ to be approximately a signed distance function, i.e. incorporating Eikonal regularization [15]. The term $\mathcal{L}_{\text{reg}}$ regularizes the residual distance Δd to be non-positive. This is because imGHUM should reside within the geometry and personalized geometric details given by Δd should be modeled on top of the skin surface.

4.2 Dynamic Semantic Human NeRF

We now extend the framework to model dynamic human motion. To implicitly model scenes where the human moves, we learn a consistent, continuous function $G_\omega : (\mathbf{c}, \alpha, \hat{d}|\mathbf{z}) \to (\mathbf{c}', \alpha', \hat{d}'|\mathbf{z}')$ for both the geometry and the appearance flow across frames. For generalization to novel poses or shapes, G_ω should be conditioned on a semantically meaningful latent code $\mathbf{z}$ that can be interpolated and should ideally have good extrapolation properties. To integrate scene information over time in a single radiance field, we follow the approaches in [36, 38] and warp observations into a canonical reference frame. Given our constraints to G_ω, imGHUM, conditioned on its semantic shape and pose latent code $(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{T})$, naturally provides spatial correspondences across frames. Given two spatial points $\mathbf{x}$ and $\mathbf{x}'$ in the space of two human instances $(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{T})$ and $(\boldsymbol{\beta}', \boldsymbol{\theta}', \mathbf{T}')$ respectively, we decide that they are in correspondence if they have the same semantics and signed distances $(\mathbf{s}, d|\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{T}) = (\mathbf{s}', d'|\boldsymbol{\beta}', \boldsymbol{\theta}', \mathbf{T}')$. Essentially, imGHUM assigns a 4D point descriptor $(\mathbf{s}, d) \in \mathbf{R}^4$ for any spatial point and deforms the volume continuously with respect to the parameterized articulated human body surface. Specifically, we apply $H_\omega : (\mathbf{T}^{-1}\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\theta}) \to (d, \mathbf{s})$ to map a spatial point $\mathbf{x}$ to a canonical point descriptor $(\mathbf{s}, d)$. We then modify the input to the NeRF function with the point descriptor as $F_w : (\mathbf{s}, d) \to (\mathbf{c}, \sigma)$, and similarly for the residual SDF function as $\Delta H_\omega : (\mathbf{s}, d) \to \Delta d$. imGHUM is pre-trained using a large corpus of 3D human scans, thus we keep it fixed in our network and directly use this prior to integrate structured geometric and appearance information across training frames. Given that the background scene is invariant to the human motion, we only apply the imGHUM warping function to spatial points inside the bounding box $\mathbf{B}$, which largely improves the computation time and memory consumption.

Pose-Dependent Geometry and Appearance. Training a single canonical NeRF and residual SDF from an image sequence integrates 2D observations into a consistent implicit geometric and volumetric radiance representation. However, we observe that fine level geometric and appearance variations caused by human motion are missing in the canonical NeRF and the residual SDF. To model such variations in geometry (clothing deformation and wrinkles) and in appearance (lighting and self-shadows), we condition the NeRF function F_ω, both the volume density σ (geometry) and color value $\mathbf{c}$ (appearance), as well as the residual SDF ΔH_ω (geometry), on the body pose code $\boldsymbol{\theta}$. In addition, we also notice that the relative position and rotation of the person with respect to the scene largely affects appearance (due to illumination effects), but should not affect the geometry. Based on this observation, we additionally condition the NeRF color on the person's root transformation as $\mathbf{c}(\mathbf{x}|\boldsymbol{\theta}, \mathbf{T})$.

H-NeRF Articulation. Differently from prior work [36, 38], all of our H-NeRF modules are now conditioned on the semantic human configuration latent code $(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{T})$, leading to an interpolatable SDF and radiance field. Moreover, due to imGHUM's capability to accurately model geometric volume deformation for diverse human poses and shapes, we have empowered H-NeRF with strong generalisation to novel poses and even body shapes. Concretely, during inference, one can simply set a different set of $(\boldsymbol{\beta}, \boldsymbol{\theta})$ to transfer the learned geometry and appearance to an unseen pose and shape configuration. For better generalization, we add Gaussian noise to the input code $(\boldsymbol{\theta}, \mathbf{T})$ during training, preventing the network to overfit to pose-dependent deformation and appearance effects seen during training. Similarly, we apply the same technique to NeRF conditioned on the view direction $\mathbf{v}$, when only a sparse set of training cameras is available.

Fine-tuning imGHUM Pose and Shape. We observe that the imGHUM fit can be a source of error when learning F_ω and ΔH_ω from dynamic image sequences. Instead of keeping them fixed, we further improve the fit during training by fine tuning a time-consistent shape correction $\Delta\boldsymbol{\beta}$ and per-frame pose correction $\Delta\boldsymbol{\theta}(t)$ with

$$\mathcal{L}_{\text{fit}} = \sum_{\mathbf{r}\in\mathcal{R}} \text{BCE}(\mathbf{M}(\mathbf{r}), d_{\text{min}}(\mathbf{r}|\boldsymbol{\beta} + \Delta\boldsymbol{\beta}, \boldsymbol{\theta} + \Delta\boldsymbol{\theta}(t))), \quad \mathcal{L}_{\text{inc}} = \|\Delta\boldsymbol{\beta}\|_2 + \frac{1}{N_t}\sum_{t=1}^{N_t} \|\Delta\boldsymbol{\theta}\|_2, \tag{11}$$

where $\mathcal{L}_{\text{fit}}$ aligns the imGHUM fit with the human foreground segmentation and $\mathcal{L}_{\text{inc}}$ regularizes the incremental corrections.

5 Experiments

In the following, we quantitatively and qualitatively evaluate H-NeRF through ablations and by comparing it with other methods. Please refer to our supplementary materials for more results and ablation experiments. We first detail our model architecture, the used datasets, and evaluation metrics.

Architecture and Training. We adapt the same network architecture for all of our experiments, where the trainable modules F_ω and ΔH_ω are composed of an eight 256-dimensional and six 128-dimensional MLP respectively, with a skip connection to the middle layer and with Swish activation [39]. As the original NeRF [30], we use 256 coarse and fine-level ray samples, for each of which we use 8- and 1-dimensional positional encoding for F_ω and ΔH_ω, respectively. We train the network using the Adam optimizer with a learning rate of 0.001 exponentially decayed by a factor

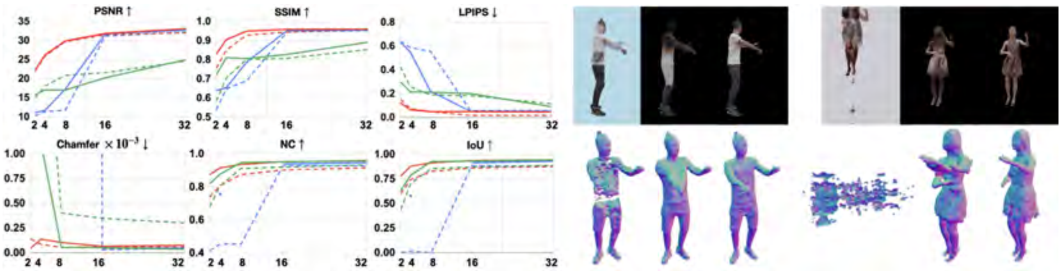

Figure 2: Left: H-NeRF (red, x-axis: #cameras) outperforms the original NeRF (blue) and IDR (green) in both novel view synthesis on 16 test cameras and 3D reconstruction of a static scene. NeRF and IDR fail under sparse camera views (solid line: RenderPeople scan; dashed line: GHS3D scan). Right: we qualitatively show novel views and reconstructed geometry for NeRF, IDR, and H-NeRF (from left to right) trained using four cameras.

Model	Dataset	PSNR ↑	SSIM ↑	LPIPS ↓	Ch $\times 10^{-3}$ ↓	NC ↑	IoU ↑
NeuralBody [37]	RenderPeople	27.33/23.52	0.888/0.827	**0.117**/0.247	0.536/0.63	0.908/0.892	0.864/0.824
	GHS3D	24.7	0.829	0.236	0.79	0.887	0.81
	PeopleSnapshot	24.62	0.849	0.160	–	–	–
	Human3.6M	24.86	0.82	0.189	–	–	–
H-NeRF (ours)	RenderPeople	**28.78/24.31**	**0.913/0.856**	0.125/**0.246**	**0.217/0.274**	**0.950/0.939**	**0.917/0.9**
	GHS3D	**24.92**	**0.852**	**0.232**	**0.218**	**0.932**	**0.89**
	PeopleSnapshot	**26.33**	**0.868**	**0.159**	–	–	–
	Human3.6M	**25.01**	**0.83**	**0.17**	–	–	–

Table 1: Quantitative evaluation of dynamic sequences on various datasets. When two metrics are reported they correspond to evaluation of training poses and novel poses at test cameras, respectively. PeopleSnapshot and Human3.6M are evaluated on novel poses under training views (with ground-truth images). Geometric metrics are only reported when ground-truth is available.

0.1 until the maximum number of iterations ($10k$ iteration of $4k$ ray batch size) is reached. We apply Gaussian noise with $\sigma = 0.1$ to the NeRF conditioning code $(\boldsymbol{\theta}, \mathbf{T}, \mathbf{v})$.

Dataset and Metrics. We provide both qualitative and quantitative evaluation on four different datasets: RenderPeople (8 sequences), GHS3D (14 sequences), PeopleSnapshot [4] (7 sequences) and Human3.6M [20] (5 sequences). The first two are synthetic datasets, rendering animated (RenderPeople) or 4D human scans (GHS3D) from four orthogonal cameras, where we have ground-truth geometry paired with images. We evaluate both image and 3D geometry reconstruction quality from two novel cameras, and qualitatively demonstrate the generalization capability to novel shapes and poses. The remaining datasets are real captured videos (PeopleSnapshot: monocular, Human3.6M: four cameras) without paired 3D geometry, where we only evaluate the rendered images. For image metrics, we adopt peak signal-to-noise ratio (PSNR ↑), structural similarity index (SSIM ↑) and learned perceptual image patch similarity (LPIPS ↓) [59]. For evaluation, we render the NeRF field inside the human bounding box $\mathbf{B}$ and compare to the ground-truth segmented image within a region of interest around the person. To evaluate geometric reconstruction quality, we report bi-directional Chamfer (Ch) L_2 distance, Normal Consistency (NC) and Volumetric Intersection over Union (IoU), evaluated on the mesh produced by Marching Cubes of the SDF with a resolution of 256^3.

Static Human Reconstruction under Sparse Cameras. H-NeRF learns a geometrically regularized volumetric radiance field unified with an imGHUM-based SDF, which significantly improves the robustness to sparse training camera views. In fig. 2, we reconstruct a static RenderPeople scan and a static GHS3D scan using the original NeRF [30], the state-of-the-art multi-view reconstruction approach IDR [54], and our H-NeRF with increasing number of cameras. Both, novel view image quality and geometry accuracy improve for all methods with increasing number of training camera images. However, in contrast to the competing methods, H-NeRF produces high quality output even under very sparse (2-8) training cameras, demonstrating the usefulness of our co-training approach.

Dynamic Human Reconstruction and Rendering. In the following, we evaluate H-NeRF's capabilities to reconstruct and render dynamic scenes (fig. 5). We compare H-NeRF against NeuralBody [37] (tab. 1), the current state-of-the-art for novel view synthesis of humans in motion. For fair comparison, we have trained NeuralBody with the same data as H-NeRF and use GHUM meshes (instead

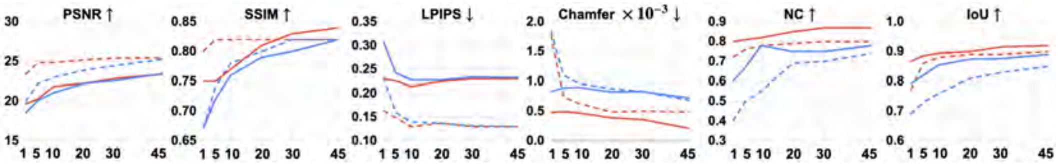

Figure 3: Frame Number Ablation. H-NeRF (red, x-axis: #training video frames) outperforms Neural Body (blue), evaluated on a RenderPeople (dashed) and a GHS3D (solid) sequence.

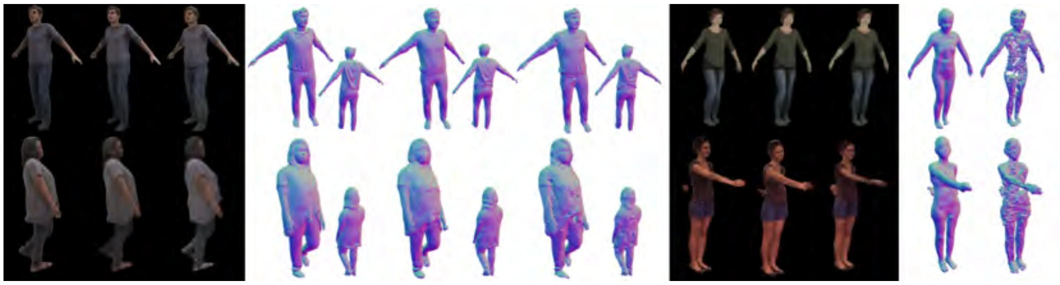

Figure 4: Qualitative results and comparisons with NeuralBody on the four different datasets. Left: GHS3D (top) and RenderPeople (bottom) ground-truth image, our result, NeuralBody, ground-truth geometry (front and back), our geometry, NeuralBody's geometry. Right: PeopleSnapshot (top) and Human3.6M (bottom) ground-truth image, our result, NeuralBody, our geometry, NeuralBody's geometry. Pay attention to the sharp renderings and the complete and detailed geometry produced by our method. Digital zoom-in recommended.

of SMPL) produced with the same pose and shape codes as used in our model for NeuralBody's structured latent codes. We further compare with Nerfies [36]. However, Nerfies has difficulties with the high degree of motion in our test sequences and fails for most of them. To this end, we consider the results as less meaningful and we report them only in the supplemental material for completeness.

H-NeRF is superior over NeuralBody across datasets and metrics, especially on geometry reconstruction. One possible explanation for NeuralBody's worse performance on view synthesis, is the fact that NeuralBody is not conditioned on the root transformation. Thus, global lighting effect are less well handled. Further, pose depended effects that go beyond body articulation are only implicitly modeled through relative distances of input mesh vertices. Finally, NeuralBody conditions on the frame index, which is set to zero for novel poses. On the PeopleSnapshot dataset where no view-depended and only subtle pose-depended effects are present, these limitations are not so apparent. In contrast, H-NeRF pays special attention to model both detailed geometry and appearance and is explicitly conditioned on pose and the root transformation for pose-depended geometry and appearance effects. This strategy pays off, especially in more complex scenarios.

Frame Number Ablation. We have now demonstrated H-NeRF's rendering and reconstruction capabilities for static and dynamic sequences. Next, we evaluate how many different poses or frames H-NeRF has to see during training for good pose extrapolation. We again compare against NeuralBody in this experiment. To this end, we have trained both methods with increasing number of frames uniformly distributed over the full sequence and compare the results in fig. 3. As expected, both methods perform better when trained with more data. However, H-NeRF performs overall better and more importantly, the quality degrades less for few frames, especially for geometry reconstruction. Our results demonstrate once more H-NeRF's robustness to little amounts of training data and its capability to reconstruct accurate geometry. To sum up, H-NeRF can be robustly trained with as little as 10 frames per camera (in a four camera set-up), resulting in a capturing effort of only minutes in practise.

Qualitative Results and Pose/Shape Extrapolation. Finally, we demonstrate H-NeRF's rendering and reconstruction accuracy also qualitatively (fig. 4). Consistent with the reported metrics, our synthesized novel poses appear sharper, more detailed, and contain less noise than current state-of-the-art. The estimated geometry is complete, smooth, and contains much of the detail present in the original scan, e.g. the clothing folds on the back of the person in the first row left. In contrast, geometries produced by NeuralBody are much more noisy and sometimes even incomplete. To

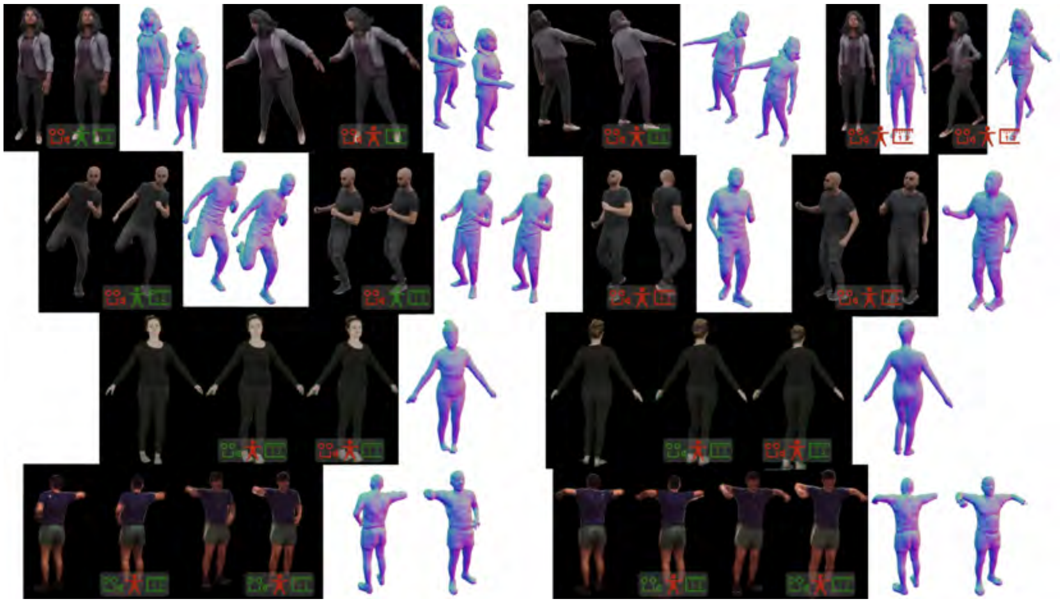

Figure 5: Qualitative results of H-NeRF on novel image synthesis. We show ground truth images and scans (left, if available) side-by-side with our results (right). Red icons correspond to test, green icons to training configuration. Symbols correspond to camera, pose, and shape, respectively. From top to bottom: RenderPeople, GHS3D, PeopleSnapshot and Human3.6M.

further demonstrate the versatility of our approach, we show examples of synthesized images and reconstructed geometry from novel viewpoints, in novel poses, and with altered body shape in fig. 5.

Limitations and Broader Impact. In the current setup, we assume full-body views of a single person in every-day clothing. Violating these assumptions, e.g. by placing another person in the scene would break the method although models of partial views or representation estimates of multiple people could be used towards more generality. Furthermore, our method has some sensitivity to the estimated body shapes and poses, as well as the quality of image segmentation. Although the process of learning to render could absorb certain inaccuracies, as pose, shape or segmentation are increasingly degraded, the method would fail eventually.

Our method paves the way towards novel immersive AR and VR applications. However, our approach is not targeted, or particularly useful for applications like visual surveillance or person identification as a particular set-up and a cooperating subject is needed. We do not build a audio-visual model that would be typical sought for deep fakes.

6 Conclusions

We have presented novel neural radiance field models for the photo-realistic rendering and the temporal reconstruction of humans in motion. Our objective is to extend the scope of the previous state of the art, focused on static scenes observed by a large number of cameras, by building dynamic models, which based on a very sparse set of views, can generalize well to novel camera views, human body poses, and even human body shapes. Our key contribution is in developing a new model with associated multiple losses, to specialize and constrain a generic NeRF formulation by using a compatible implicit statistical 3D human pose and shape model represented using signed distance functions. Our implicit geometric formulation captures not just the statistical human body form but also hair and clothing represented as an implicit residual network. Our model is trained end-to-end based on several novel losses and achieves very good results for both 3D reconstruction and rendering. Training the body model in an end-to-end rendering framework carries the promise to learn complex implicit skinning functions based on images only, a performance previously possible only in the realm of human capture using complex, laboratory placed, expensive 3D body scanners. We illustrate the favorable capabilities of our model by extensive experimentation using several datasets and against other state of the art techniques.

Supplementary Material

In this supplementary material we provide additional results, both quantitative and qualitative, as well as comparisons with other methods, ablation studies, and additional background on our model components.

A Additional Results

Dynamic Human Reconstruction and Rendering. In fig. 6, we provide qualitative results for samples taken from the four datasets we experiment with. Our method shows very good accuracy for both novel view synthesis and for 3D geometric reconstruction. In addition, our model allows to generate novel geometries corresponding to poses not in the training set, by altering the input imGHUM pose control $(\boldsymbol{\theta}, \mathbf{T})$, as shown in the fig. 6. We note that differently from novel view synthesis of a fixed pose, the source of error in the image synthesis of a novel pose (not in the training set) could come from both the NeRF rendering function and the implicit geometric surface skinning. For our RenderPeople sequences where ground-truth images (renderings) and geometries for the novel test poses are available, we quantitatively report both image and geometric errors in tab. 1 of the main paper. We also show side-by-side qualitative comparisons in fig. 4 (main paper) and fig. 6 of this material.

Tab. 2 provides additional information on the used datasets. Specifically, the RenderPeople sequences are generated by animating a single rigged scan using motion capture from the CMU [1] and Human3.6M [20] datasets, respectively. The GHS3D dataset is composed of 14 sequences dynamically capturing dressed subjects undertaking freestyle motions (e.g. presentation, dancing or exercising). The PeopleSnapshot [4] dataset coonsists of monocular videos of a subject rotating in front of a static camera. We select one rotation cycle for training and another (different) cycle for testing. We also demonstrate the capabilities of our method on the Human3.6M dataset which captures the scene using four synchronized and calibrated cameras. There we use the first 320 frames for training (sampled every 8th frame) and the following 160 frames for testing (sampled every 4th frame).

Comparison to Nerfies [36]. For large full-body articulated human motion, as in our typical use case (and illustrated by our sequences), Nerfie overfits to the training images by mixing the human geometry with the background. This results in extremely limited generalisation for both novel view synthesis and 3D geometric reconstruction (tab. 3). We notice that geometry is effectively not reconstructed, leading to zero IoU. Moreover, similarly to D-NeRF [38], the deformation function is conditioned on time (or an embedding of the video frame index), hence the method cannot generalize to novel poses or shapes. To evaluate image (rendering) quality, we crop the image using a 2D bounding box computed from the ground-truth image segmentation (padded with a 20 pixel border), as our region of interest. For H-NeRF, IDR [54] and NeuralBody [37], the evaluation is performed by comparing to the cropped ground-truth image and using the foreground human segmentation. However, for the original NeRF and Nerfies, segmentation is not produced and therefore we compute metrics on the image with background.

Generalization. In addition to pose generalization shown in fig. 6, we further evaluate novel shape generalization in fig. 7 for sample sequences, taken from the four datasets. The shape extrapolation is achieved by replacing the training imGHUM shape latent code with a different $\boldsymbol{\beta}$, and propagate the shape changes to both image rendering and the geometry. To understand this process, one can consider imGHUM as the inner layer of the human body. The color and signed distance of a spatial point is deforming w.r.t. the body surface. Therefore when we change the underlying body shape, the NeRF radiance field and the signed distance function are updated accordingly. We observe high-quality visual shape generalization results produced by H-NeRF, in fig. 7, even for significant volume changes w.r.t. the ground truth shape.

In fig. 8, we study the correlation of the image synthesis quality w.r.t. differences between testing and training camera views. The view difference (x-axis) is evaluated by computing the angle between the two vectors from the camera position to the 3D center of the person. All methods (H-NeRF, IDR and NeRF) show a degradation of quality as viewpoints deviate significantly from training, but the view synthesis capability of H-NeRF is consistently better than the other two, likely given its capacity to estimate a reasonable 3D surface geometry. The increased error is largely due to the rendering of body parts not visible in the training images. We note that for our RenderPeople and GHS3D (static

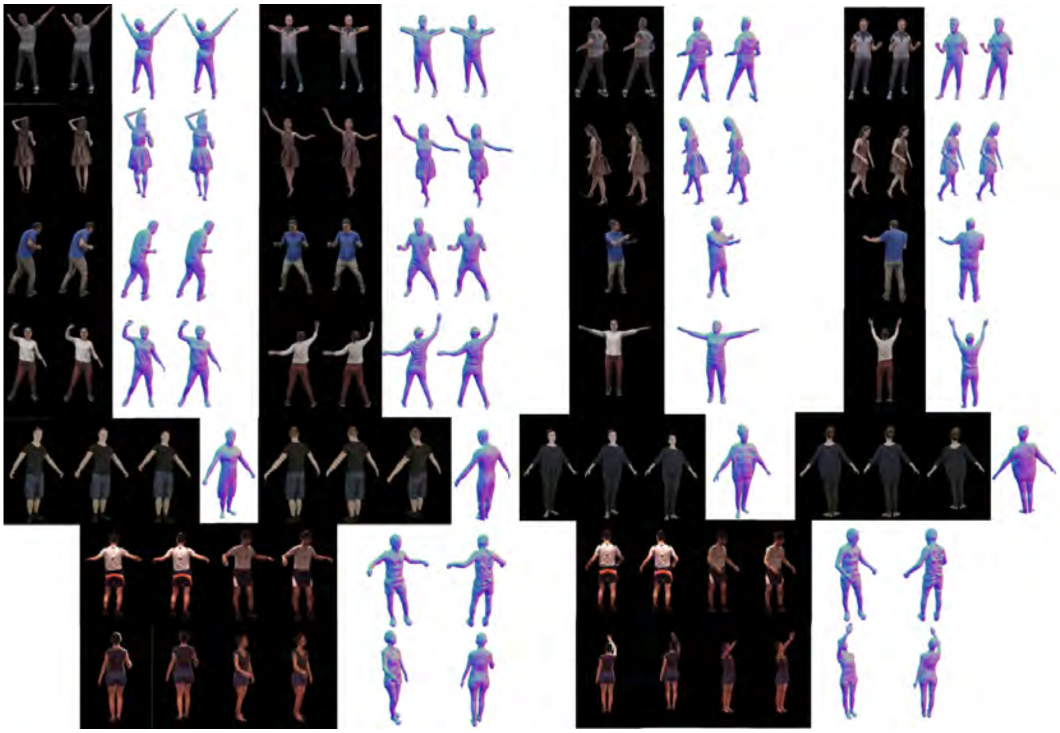

Figure 6: Qualitative results for dynamic sequences. When available, we show ground-truth (left) image and geometry side by side with our results (right). The top 2 rows show single scan animations based on RenderPeople assets, showing novel view synthesis (the left 2) and novel pose generalization (the right 2, under novel views). Similarly, the 3rd-4th row show GHS3D scan sequences, with 2 novel view synthesis and 2 pose generalization illustrations. The 5th row shows the monocular PeopleSnapshot [4] ground-truth image, our rendering of a novel pose under the learned and a novel camera, respectively, as well as our geometric reconstruction from left to the right. The last 2 rows illustrate the pose generalization capability of our models on Human3.6M [20], where we render the novel pose under 2 training camera views and show ground-truth images for side-by-side comparison.

Train

Dataset	#seq	#cam	#poses	scene-scale	gt-image	gt-geom	opt
RenderPeople	8	4	39/53/71	2.5	Yes	Yes	No
GHS3D	14	4	31/44/59	2.5	Yes	Yes	Yes
PeopleSnapshot	9	1	19/29/55	2.5	Yes	No	Yes
Human3.6M	5	4	40/40/40	5	Yes	No	Yes

Test

Dataset	#seq	#cam	#poses	gt-image	gt-geom
RenderPeople	8	2	32/47/68	Yes	Yes
GHS3D	14	2	40/47/59	No	No
PeopleSnapshot	9	1	22/31/50	Yes	No
Human3.6M	5	4	40/40/40	Yes	No

Table 2: Dataset statistics. #poses are reported as the minimal/median/maximum number of poses. Scene-scale represents the size of the scene where we divide it to scale into $[-1,-1,-1]$ to $[1,1,1]$. gt-image and gt-geom indicate that ground-truth image and geometry are available, respectively. If opt is yes, we perform fine-tuning of imGHUM during training.

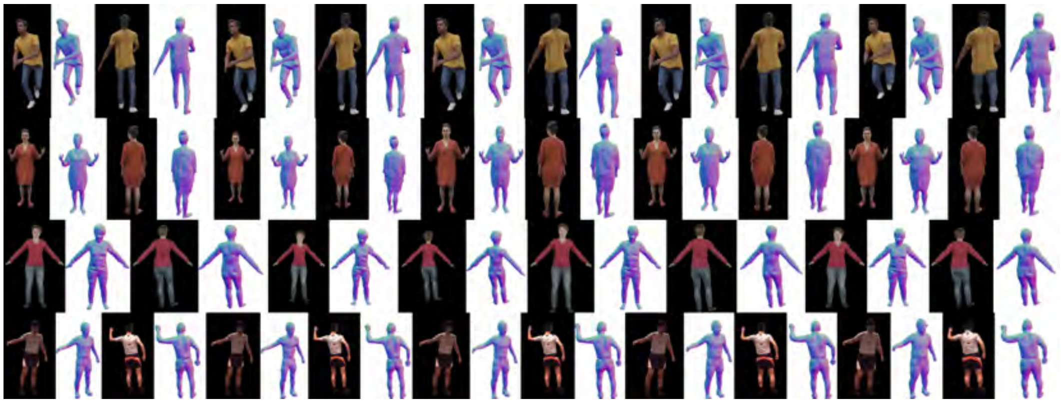

Figure 7: Qualitative evaluation of shape extrapolation for the four datasets. The leftmost is the training shape and we extrapolate both the rendering and the geometric reconstruction to four novel shapes (three for PeopleSnapshot). The shape volume change w.r.t. the learned shape for the 4 rows are: $-5\%, +29\%, +50\%, +87\%$; $-27\%, +48\%, +18\%, +32\%$; $-41\%, +18\%, +9\%$; $-15\%, +4\%, +18\%, +44\%$.

or dynamic) sequences, the four training cameras are on the four sides of the subject, residing on a sphere with radius 2.4m, oriented towards the center as shown in the last row of fig. 8, where we also show two typical test camera views for our dynamic sequences.

We further evaluate the pose generalization capability of H-NeRF w.r.t. the pose difference to the training configurations. Fig. 9 shows the image and geometric metrics of two RenderPeople sequences for all of our novel testing poses. We do not observe strong correlation as a function of pose differences w.r.t. the training. and suspect that both image and geometric quality are largely affected by the accuracy of imGHUM on the test poses, and less so by pose differences. However, one still needs to provide sufficient examples for learning pose-dependent geometry and appearance as shown in our frame number ablation experiment (fig.3 in the main paper).

Monocular videos. The PeopleSnapshot dataset consists of monocular videos of people rotating in full circle in front of the camera. However, the subjects are consistently holding an A-Pose. We therefore additionally evaluate H-NeRF on two RenderPeople and two GHS3D dynamic sequences with large pose variations but trained with only a single camera. Tab. 4 and fig. 10, respectively, show numerical and visual results. Our method still shows strong robustness for novel view synthesis, geometric reconstruction and even pose extrapolation, although at some quality expense compared to models trained using four cameras. We notice that GHS3D-S47 shows a more significant drop in its metrics when trained in the monocular case, which may be caused by the back of the subject not being fully visible in the training images (whereas for the other sequences, the subject is turning in front of the camera). This experiment clearly demonstrates the capability of H-NeRF to integrate temporal image information into consistent neural radiance and SDF representations.

Similar to NeuralBody [37], we rely on good pose estimation for consistent integration of 2D observations across multiple dynamic frames. Because of degraded monocular pose estimation, we excluded a small number of side view frames from the PeopleSnapshot dataset aiming to maximize the model quality. Note that for fair comparisons, we trained NeuralBody with the same images and poses as our method. In addition, we have trained a RenderPeople sequence with all frames and have indeed observed degraded novel image synthesis quality caused by the pose estimation inaccuracies (PSNR ↑: 27.5 (full) vs. 28.1 (filtered), SSIM ↑: 0.81 vs. 0.86, LPIPS ↓: 0.23 vs. 0.21). Our process of fine-tuning imGHUM pose and shape (Eq. 11 in the main paper) helps alleviate the problem (see Tab 5) and any other better pose estimation technique would be complementary to our approach.

B Ablation Studies

Tab. 5 shows ablation studies for each loss for our four datasets. Training H-NeRF with all proposed losses strikes a good balance between image rendering quality and geometric reconstruction accuracy, and achieve the best performance for most metrics. For example, we observe significant drops in image quality when we do not condition the NeRF appearance under the root transformation. We largely benefit from fine-tuning the imGHUM parameters for the PeopleSnapshot dataset where we have the highest geometric fitting error given the use of only monocular videos. The terms $\mathcal{L}_{\text{blend}}$ and $\mathcal{L}_{\text{mask}}$ impact the image metrics significantly. For example, without $\mathcal{L}_{blend}$, the geometric metric is slightly better but leads to significantly worse SSIM for the real-world videos. The terms $\mathcal{L}_{\text{geom}}$, $\mathcal{L}_{\text{reg}}$ and $\mathcal{L}_{\text{eik}}$ affect the geometric metrics more, and we observe much worse reconstruction performance when turning these off. The Eikonal regularization helps smoothing the surface reconstruction and is more critical for videos with insufficient views (PeopleSnapshot) or lower image resolution e.g. due to people placed farther away from the camera (Human3.6M). The term $\mathcal{L}_{\text{seg}}$ affects both rendering and geometry and we clearly see performance drops in its absence.

C Training, Memory Consumption and Timings

We have trained the models for 10k iterations of 4k ray batch size on 8 Nvidia v100 GPUs, which takes about 6-8 hours for each dynamic sequence. For an image of 512×512 resolution, the inference for H-NeRF on a single Nvidia v100 GPU takes about 9.1 sec. The main computation overhead compared to the original NeRF formulation (about 6.5 sec) comes from the imGHUM warping. imGHUM warping also limits the maximum number of query points to be 64^3 due to memory constraints (i.e. 1024 rays consisting of 128 coarse and 128 fine samples). We utilize the coarse scene structuring such that we only query imGHUM for points inside the 3D bounding box $\mathbf{B}$. For exterior points (static background or free space) we use the original position coordinate and a constant distance value (1.0) as input features to the NeRF network. In practice, this significantly reduces the number of imGHUM query points by 90% and largely alleviates the memory constraints.

Parameters. For our training, the weights of $\mathcal{L}_{\text{rec}}, \mathcal{L}_{\text{mask}}, \mathcal{L}_{\text{blend}}, \mathcal{L}_{\text{geom}}, \mathcal{L}_{\text{seg}}, \mathcal{L}_{\text{reg}}, \mathcal{L}_{\text{eik}}, \mathcal{L}_{\text{fit}}$ and $\mathcal{L}_{\text{inc}}$ are validated to $1.0, 1.0, 0.1, 0.02, 0.02, 0.0005, 10^{-6}, 1.0$ and 0.02 respectively. We set $\sigma_h = 50$, $\gamma = 200$, and $\eta = 1.0$ for RenderPeople and GHS3D sequences but $\eta = 0.25$ for the other two datasets.

D imGHUM Fitting

imGHUM shares the same shape and pose latent code with the mesh-based generative human model GHUM [52]. To fit the imGHUM parameters to a given image, we therefore use an off-the-shelf fitting approach for GHUM [56, 57]. We subsequently use the GHUM fit (pose and shape) in order to initialize imGHUM, which shares the same latent code. Specifically, given a monocular image or a set of images collected from multiple views, we use the neural network HUND [57] that takes a cropped human detection, and outputs 137 keypoints with confidences and 15 body-part semantic segmentation masks. HUND further predicts the GHUM shape $\boldsymbol{\beta}$ and pose values $\boldsymbol{\theta}$ based on the various semantic features (keypoints and segmentation masks) extracted from the image. Given HUND predictions, we follow up with a kinematic optimization that better aligns GHUM's predicted keypoints and semantics segmentation with corresponding primitives extracted in the calibrated input cameras. The optimization is formulated with image alignment for which we adopt the self-supervised keypoint and body segmentation loss from [56] (cf. their eq. 11), with L_2 regularization on the latent embeddings of the body shape $\boldsymbol{\beta}$ and pose $\boldsymbol{\theta}$. The temporal smoothness is applied with a L_2 loss on the pose differences between neighboring frames.

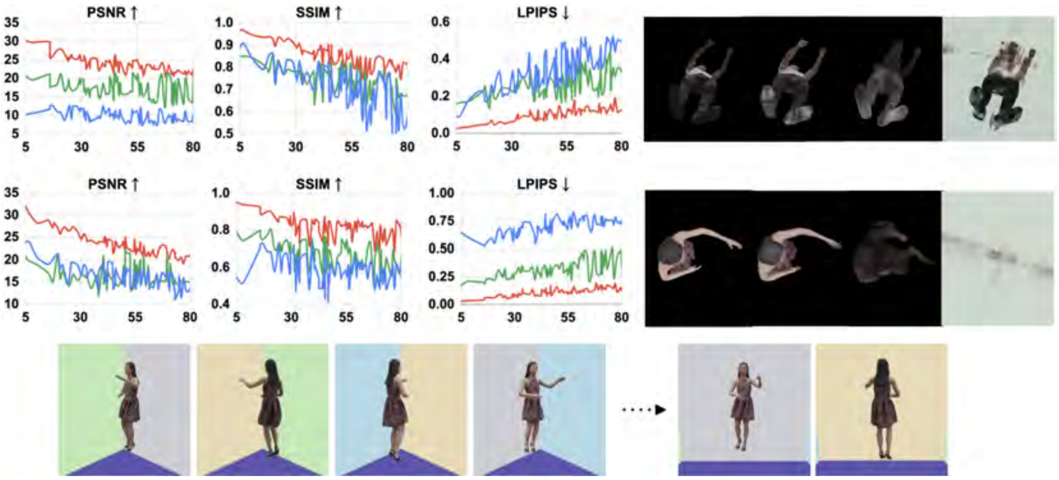

Figure 8: Image evaluation of novel view synthesis w.r.t. the minimal view difference to the training cameras (x-axis, in degrees). We evaluate on two static scenes with a model trained using four cameras. All models degrade in view synthesis quality when the test camera deviates from the training views but H-NeRF (red) significantly outperforms NeRF (blue) and IDR (green). Right: we show the ground truth, H-NeRF, IDR and NeRF from left to right, rendered under significantly different views compared to the training cameras. We note that H-NeRF still produces reasonable rendering, with image metrics degrading mostly due to body parts (the elbow pit, bottom of the shoes, etc) not visible in the training images. The last row shows the four training views on the left, and our two default test views rotated by 45 degrees away from training views.

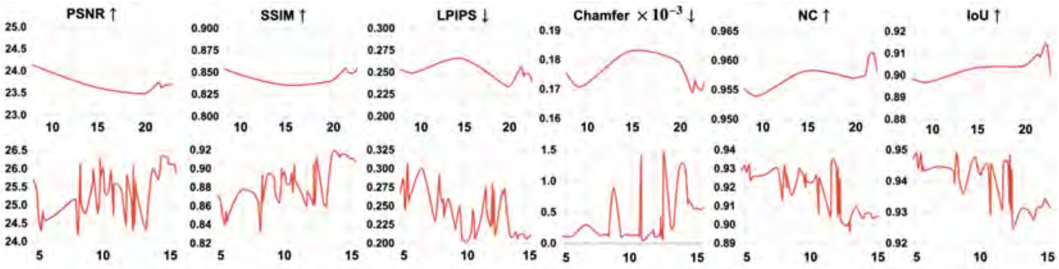

Figure 9: Image and geometric evaluation of H-NeRF on novel poses w.r.t. pose difference to the closest training pose (x-axis, in degrees, per body joint). The plots are based on two of our RenderPeople sequences. We do not observe significant degradation of the reconstruction quality with pose differences, as long as imGHUM produces good predictions for the test poses.

Model	Dataset	PSNR ↑	SSIM ↑	LPIPS ↓	Ch $\times 10^{-3}$ ↓	NC ↑	IoU ↑
Nerfie [36]	RenderPeople	10.97	0.7	0.642	57.6	0.487	0
	GHS3D	10.8	0.685	0.635	50.1	0.49	0
H-NeRF (ours)	RenderPeople	**28.78**	**0.913**	**0.125**	**0.217**	**0.950**	**0.917**
	GHS3D	**24.92**	**0.852**	**0.232**	**0.218**	**0.932**	**0.89**

Table 3: Quantitative comparisons to Nerfie [36] on our dynamic sequences. Geometric metrics are only reported when ground-truth is available. We note that for Nerfie, Marching Cubes fails in geometric reconstruction for 5/8 RenderPeople and 13/14 GHS3D sequences. Hence, we only report numbers based on the the scenes where Nerfie produces outputs. Nerfie does not support novel poses, therefore all metrics are evaluated on training poses. We do not report metrics for PeopleSnapshot and Human3.6M since no novel camera view with ground-truth images is available.

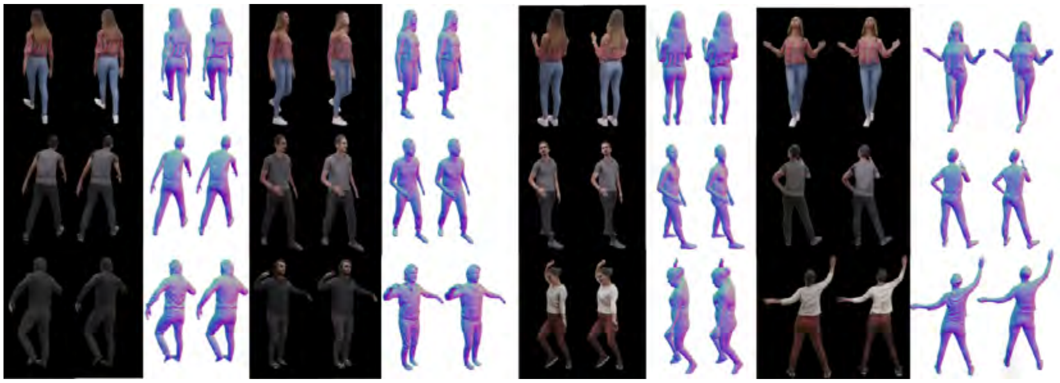

Figure 10: Qualitative evaluation of four dynamic monocular sequences. The first two rows correspond to RenderPeople-Tina and RenderPeople-Nagy, showing two novel view synthesis results for the training poses (on the left) and two novel poses seen from novel viewpoints (ground-truth image and geometry are shown by the left side). The last row shows novel view synthesis for two GHS3D sequences (S36 and S47 respectively).

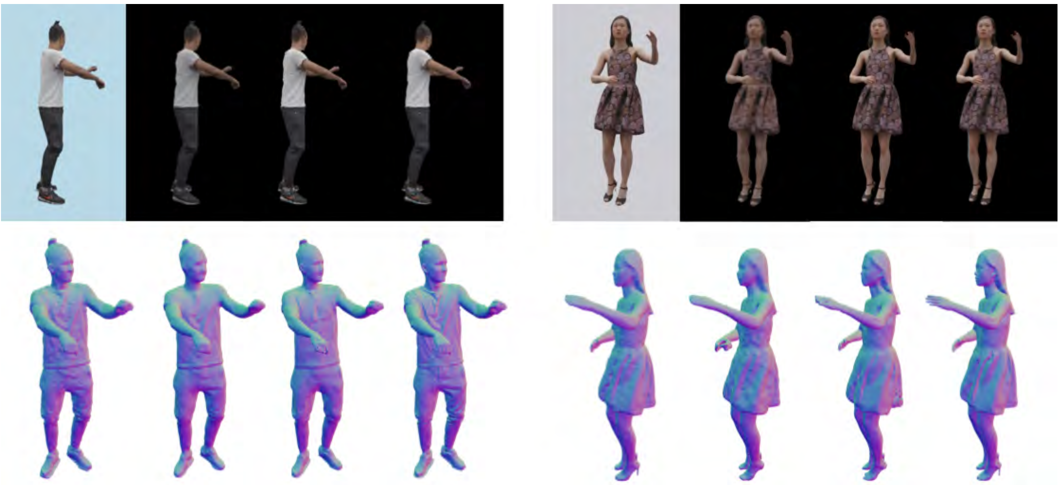

Figure 11: Qualitative evaluation of the static scenes with 32 cameras (from left to right: NeRF, IDR, H-NeRF and ground truth). With growing number of training cameras, the novel view synthesis and geometric reconstruction improve for all methods. NeRF and H-NeRF starts to converge whereas the IDR shows some differences due to the different image formation process.

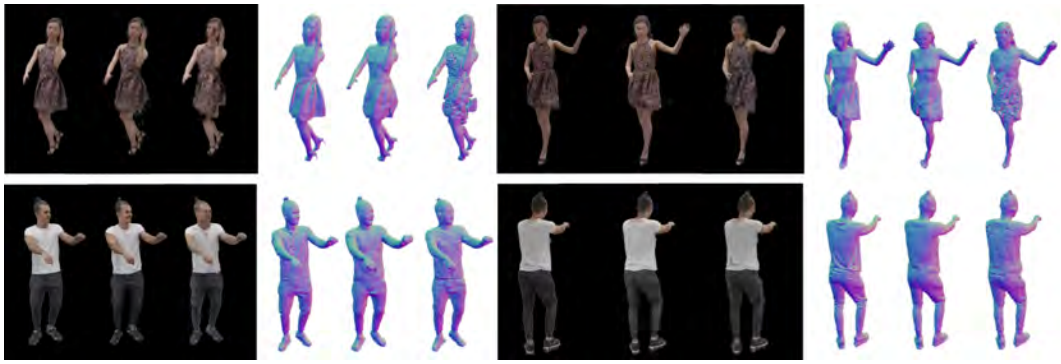

Figure 12: Qualitative comparison of H-NeRF and NeuralBody trained with sparse video frames (from left to right: ground truth, H-NeRF and NeuralBody). We visualize the novel view synthesis and geometric reconstruction for a RenderPeople (trained with 10 frames) and a GHS3D sequence (trained with 20 frames). Our qualitative results (H-NeRF) consistently outperform NeuralBody.

Sequence	#cam	PSNR ↑	SSIM ↑	LPIPS ↓
RenderPeople-Tina	1	24.74/23.48/21.79	0.863/0.850/0.819	0.142/0.178/0.224
	4	27.23/23.56/23.53	0.9013/0.855/0.854	0.146/0.185/0.219
RenderPeople-Nagy	1	26.41/23.81/23.58	0.882/0.843/0.839	0.128/0.263/0.312
	4	30.44/23.73/24.16	0.936/0.838/0.849	0.112/0.286/0.317
GHS3D-S36	1	28.83	0.874	0.141
	4	29.41	0.896	0.134
GHS3D-S47	1	23.9	0.852	0.153
	4	24.2	0.871	0.117

Sequence	#cam	Ch $\times 10^{-3}$ ↓	NC ↑	IoU ↑
RenderPeople-Tina	1	0.327/0.419	0.94/0.922	0.898/0.86
	4	0.274/0.3	0.95/0.935	0.918/0.886
RenderPeople-Nagy	1	0.339/0.477	0.96/0.938	0.944/0.93
	4	0.119/0.143	0.97/0.95	0.957/0.95
GHS3D-S36	1	0.124	0.935	0.8853
	4	0.079	0.947	0.916
GHS3D-S47	1	0.434	0.907	0.8
	4	0.118	0.948	0.905

Table 4: Quantitative evaluation on monocular RenderPeople and GHS3D sequences, compared to models trained using four cameras. For RenderPeople, the three image metrics are reported corresponding to rendering of the training poses under novel cameras, novel poses under the training camera, and novel poses under novel cameras, respectively, whereas the geometric metrics are reported as training poses/novel poses. For GHS3D, we only have ground-truth images and scans for the training poses and therefore we report the numbers for training poses under novel camera views.

Metric	Dataset	Full	$-\mathcal{L}_{\text{seg}}$	$-\mathcal{L}_{\text{mask}}$	$-\mathcal{L}_{\text{blend}}$	$-\mathcal{L}_{\text{geom}}$	$-\mathcal{L}_{\text{reg}}$	$-\mathcal{L}_{\text{eik}}$	-opt	-$\mathbf{T}$	-noise
PSNR ↑	RenderPeople	**28.35**	26.46	23.96	27.85	26.67	26.55	26.68	**28.35**	24.43	26.67
	GHS3D	**27.26**	26.74	25.63	26.88	26.79	25.68	26.72	26.70	25.72	26.90
	PeopleSn.	**28.12**	27.04	27.83	27.48	**28.12**	28.07	28.02	25.69	28.1	28.04
	Human3.6M	**24.4**	24.3	15.7	24.1	24.34	24.3	24.3	24.3	24.2	24.2
SSIM ↑	RenderPeople	**0.908**	0.874	0.819	0.872	0.882	0.881	0.878	**0.908**	0.867	0.879
	GHS3D	0.887	0.887	0.876	0.885	0.887	0.877	0.886	0.888	0.875	**0.891**
	PeopleSn.	**0.86**	0.84	0.835	0.478	0.858	0.858	0.858	0.329	0.832	0.857
	Human3.6M	0.874	0.875	0.69	0.651	0.871	0.87	0.874	**0.877**	0.865	0.869
LPSIS ↓	RenderPeople	**0.101**	0.115	0.152	0.106	0.113	0.115	0.115	**0.101**	0.125	0.119
	GHS3D	**0.203**	0.208	0.248	0.205	0.205	0.218	0.209	0.211	0.212	0.210
	PeopleSn.	0.206	0.228	0.205	0.211	0.196	0.196	**0.194**	0.235	0.206	0.196
	Human3.6M	**0.129**	0.133	0.671	0.138	0.13	0.132	0.134	0.135	0.134	0.132
Ch ↓	RenderPeople	0.121	0.129	0.107	**0.082**	0.128	0.136	0.103	0.121	0.1	0.107
	GHS3D	0.113	0.13	0.136	0.107	0.185	0.151	**0.104**	0.235	0.123	0.128
NC ↑	RenderPeople	**0.961**	0.958	0.958	**0.961**	0.947	0.958	0.959	**0.961**	0.96	0.96
	GHS3D	**0.947**	0.946	0.944	0.946	0.928	0.943	0.946	0.943	0.944	**0.947**
IoU ↑	RenderPeople	**0.943**	0.939	0.923	0.937	0.845	0.941	0.939	**0.943**	0.941	0.939
	GHS3D	0.915	0.913	0.914	**0.923**	0.851	0.911	0.915	0.91	0.91	0.909

Table 5: Ablation study on losses (Chamfer $\times 10^{-3}$). The last three columns (-opt, -$\mathbf{T}$, -noise) show results without imGHUM fine tuning, not conditioning the NeRF appearance using a root transformation $\mathbf{T}$, and not applying Gaussian noise to NeRF's condition code $(\boldsymbol{\theta}, \mathbf{T}, \mathbf{v})$, respectively. We do not apply imGHUM parameters fine-tuning on the RenderPeople sequence.

References

[1] CMU graphics lab motion capture database. 2009. `http://mocap.cs.cmu.edu/`.

[2] Kara-Ali Aliev, Dmitry Ulyanov, and Victor Lempitsky. Neural point-based graphics. *arXiv preprint arXiv:1906.08240*.

[3] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1175–1186. IEEE, 2019.

[4] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3D people models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 8387–8397. IEEE, 2018.

[5] Thiemo Alldieck, Hongyi Xu, and Cristian Sminchisescu. imGHUM: Implicit generative models of 3D human shape and articulated pose. In *Int. Conf. Comput. Vis.*, 2021.

[6] Federica Bogo, Michael J. Black, Matthew Loper, and Javier Romero. Detailed full-body reconstructions of moving people from monocular RGB-D sequences. In *IEEE International Conference on Computer Vision*, pages 2300–2308. IEEE, 2015.

[7] Joel Carranza, Christian Theobalt, Marcus A Magnor, and Hans-Peter Seidel. Free-viewpoint video of human actors. *ACM Trans. Graph.*, 22(3):569–577, 2003.

[8] Shenchang Eric Chen and Lance Williams. View interpolation for image synthesis. In *Proceedings of the 20th annual conference on Computer graphics and interactive techniques*, pages 279–288, 1993.

[9] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5939–5948, 2019.

[10] Edilson De Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. Performance capture from sparse multi-view video. In *ACM Trans. Graph.*, pages 1–10. 2008.

[11] Paul E Debevec, Camillo J Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 11–20, 1996.

[12] Boyang Deng, JP Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Neural articulated shape approximation. In *Eur. Conf. Comput. Vis.* Springer, August 2020.

[13] Juergen Gall, Carsten Stoll, Edilson De Aguiar, Christian Theobalt, Bodo Rosenhahn, and Hans-Peter Seidel. Motion capture using joint skeleton tracking and surface estimation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1746–1753. IEEE, 2009.

[14] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 770–785, 2018.

[15] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *Int. Conf. on Mach. Learn.*, pages 3569–3579. 2020.

[16] Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Real-time deep dynamic characters. *ACM Trans. Graph.*, 40(4), aug 2021.

[17] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *IEEE Conf. Comput. Vis. Pattern Recog.* IEEE, jun 2020.

[18] Marc Habermann, Weipeng Xu, Michael Zollhöfer, Gerard Pons-Moll, and Christian Theobalt. Livecap: Real-time human performance capture from monocular video. *ACM Trans. Graph.*, 38(2):14:1–14:17, 2019.

[19] Peter Hedman, Pratul P Srinivasan, Ben Mildenhall, Jonathan T Barron, and Paul Debevec. Baking neural radiance fields for real-time view synthesis. In *Int. Conf. Comput. Vis.*, 2021.

[20] 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.

[21] Thomas Lewiner, Hélio Lopes, Antônio Wilson Vieira, and Geovan Tavares. Efficient implementation of marching cubes' cases with topological guarantees. *Journal of Graphics Tools*, 8(2):1–15, 2003.

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

[23] Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. Deep appearance models for face rendering. *ACM Trans. Graph.*, 37(4):1–13, 2018.

[24] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *ACM Trans. Graph.*, 38(4):65:1–65:14, July 2019.

[25] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16, 2015.

[26] Lars M. Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4460–4470, 2019.

[27] Moustafa Meshry, Dan B Goldman, Sameh Khamis, Hugues Hoppe, Rohit Pandey, Noah Snavely, and Ricardo Martin-Brualla. Neural rerendering in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6878–6887, 2019.

[28] Mateusz Michalkiewicz, Jhony K Pontes, Dominic Jack, Mahsa Baktashmotlagh, and Anders Eriksson. Deep level sets: Implicit surface representations for 3D shape inference. *arXiv preprint arXiv:1901.06802*, 2019.

[29] Marko Mihajlovic, Yan Zhang, Michael J. Black, and Siyu Tang. LEAP: Learning articulated occupancy of people. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

[30] 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 *Eur. Conf. Comput. Vis.*, 2020.

[31] Roger Mohr, Long Quan, and Françoise Veillon. Relative 3d reconstruction using multiple uncalibrated images. *The International Journal of Robotics Research*, 14(6):619–632, 1995.

[32] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *IEEE Conf. Comput. Vis. Pattern Recog.*, June 2020.

[33] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *Int. Conf. Comput. Vis.*, 2021.

[34] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Int. Conf. Comput. Vis.*, 2021.

[35] Jeong Joon Park, Peter Florence, Julian Straub, Richard A. Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 165–174, 2019.

[36] 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 *Int. Conf. Comput. Vis.*, 2021.

[37] 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 *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

[38] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *IEEE Conf. Comput. Vis. Pattern Recog.* IEEE, jun 2021.

[39] Prajit Ramachandran, Barret Zoph, and Quoc V Le. Searching for activation functions. *arXiv preprint arXiv:1710.05941*, 2017.

[40] Daniel Rebain, Wei Jiang, Soroosh Yazdani, Ke Li, Kwang Moo Yi, and Andrea Tagliasacchi. Derf: Decomposed radiance fields. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

[41] Gernot Riegler and Vladlen Koltun. Free view synthesis. In *European Conference on Computer Vision*, pages 623–640. Springer, 2020.

[42] 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 *Int. Conf. Comput. Vis.*, pages 2304–2314, 2019.

[43] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016.

[44] Aliaksandra Shysheya, Egor Zakharov, Kara-Ali Aliev, Renat Bashirov, Egor Burkov, Karim Iskakov, Aleksei Ivakhnenko, Yury Malkov, Igor Pasechnik, Dmitry Ulyanov, et al. Textured neural avatars. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2387–2397, 2019.

[45] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Adv. Neural Inform. Process. Syst.*, 33, 2020.

[46] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhöfer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2019.

[47] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Adv. Neural Inform. Process. Syst.*, 2019.

[48] Shih-Yang Su, Frank Yu, Michael Zollhoefer, and Helge Rhodin. A-nerf: Surface-free human 3d pose refinement via neural rendering. In *Adv. Neural Inform. Process. Syst.*, 2021.

[49] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, et al. State of the art on neural rendering. In *Comput. Graph. Forum*, volume 39, pages 701–727. Wiley Online Library, 2020.

[50] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019.

[51] Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović. Articulated mesh animation from multi-view silhouettes. In *ACM Trans. Graph.*, pages 1–9. 2008.

[52] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3d human shape and articulated pose models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6184–6193, 2021.

[53] Weipeng Xu, Avishek Chatterjee, Michael Zollhoefer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. Monoperfcap: Human performance capture from monocular video. *ACM Trans. Graph.*, 2018.

[54] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Adv. Neural Inform. Process. Syst.*, 33, 2020.

[55] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Int. Conf. Comput. Vis.*, 2021.

[56] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. In *Eur. Conf. Comput. Vis.*, 2020.

[57] Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Andrei Zanfir, and Cristian Sminchisescu. Human synthesis and scene compositing. In *AAAI*, pages 12749–12756, 2020.

[58] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020.

[59] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.