Learning Neural Volumetric Representations of Dynamic Humans in Minutes

Chen Geng* Sida Peng* Zhen Xu* Hujun Bao Xiaowei Zhou†

State Key Laboratory of CAD&CG, Zhejiang University

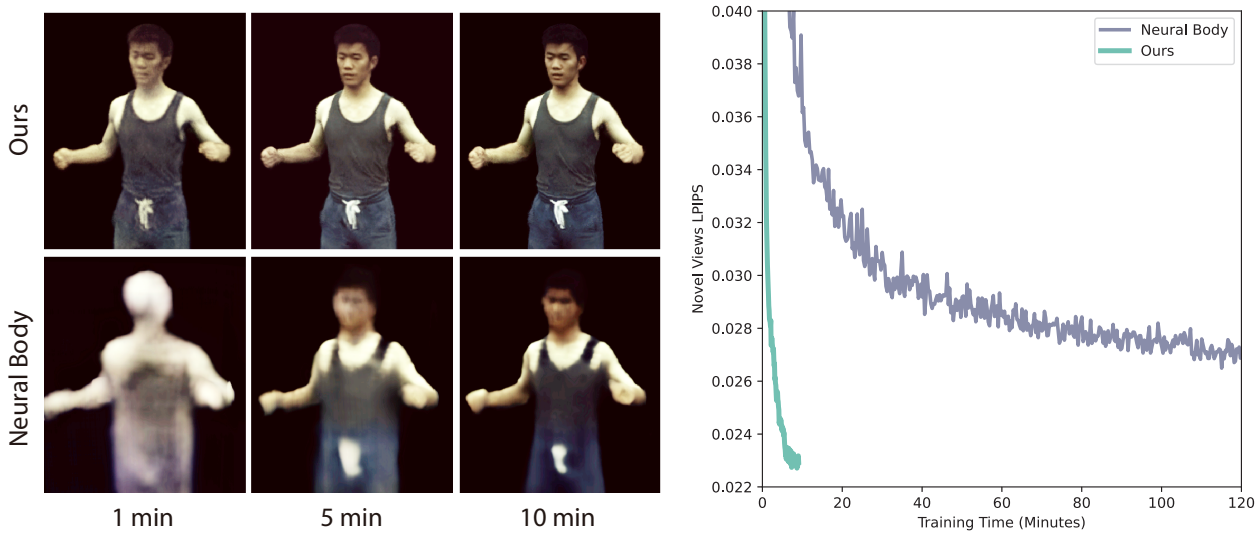

Figure 1. **Convergence rate of training.** Given a monocular video of a human performer, our model can be learned in ∼5 minutes to produce photorealistic novel view rendering, which is 100 times faster than Neural Body [58].

Abstract

This paper addresses the challenge of efficiently reconstructing volumetric videos of dynamic humans from sparse multi-view videos. Some recent works represent a dynamic human as a canonical neural radiance field (NeRF) and a motion field, which are learned from input videos through differentiable rendering. But the per-scene optimization generally requires hours. Other generalizable NeRF models leverage learned prior from datasets to reduce the optimization time by only finetuning on new scenes at the cost of visual fidelity. In this paper, we propose a novel method for learning neural volumetric representations of dynamic humans in minutes with competitive visual quality. Specifically, we define a novel part-based voxelized human representation to better distribute the representational power of the network to different human parts. Furthermore, we propose a novel 2D motion parameterization scheme to increase the convergence rate of deformation field learning. Experiments demonstrate that our model can be learned 100 times faster than previous per-scene optimization methods while being competitive in the rendering quality. Training our model on a 512×512 *video with 100 frames typically takes about 5 minutes on a single RTX 3090 GPU. The code is available on our project page: https://zju3dv.github.io/instant_nvr.*

1. Introduction

Creating volumetric videos of human performers has many applications, such as immersive telepresence, video games, and movie production. Recently, some methods [58, 93] have shown that high-quality volumetric videos can be recovered from sparse multi-view videos by representing dynamic humans with neural scene representations. However, they typically require more than 10 hours of training on

*Equal contribution. †Corresponding author.

a single GPU. The expensive time and computational costs limit the large-scale application of volumetric videos. Generalizable methods [34, 100] utilize learned prior from datasets of dynamic humans to reduce the training time by only finetuning on novel human performers. These techniques could increase the optimization speed by a factor of 2-5 at the cost of some visual fidelity.

To speed up the process of optimizing a neural representation for view synthesis of dynamic humans, we analyze the structural prior of the human body and motion, and propose a novel dynamic human representation that achieves 100x speedup during optimization while maintaining competitive visual fidelity. Specifically, to model a dynamic human, we first transform world-space points to a canonical space using a novel motion parameterization scheme and inverse linear blend skinning (LBS) [35]. Then, the color and density of these points are estimated using the canonical human model.

The innovation of our proposed representation is two-fold. First, we observe that human body parts have different levels of complexity in terms of both shape and texture. For example, the face of a human performer typically exhibits higher complexity than a flatly textured torso region, thus requiring more representational power to depict. Motivated by this, our method decomposes the canonical human body into multiple parts and represents the human body with a structured set of voxelized NeRF [47] networks to bring the convergence rate of these different parts to the same level. In contrast to a single-resolution representation, the part-based body model utilizes the human body prior to represent the human shape and texture efficiently, heuristically distributing variational representational power to human parts with spatially varying complexity.

Second, we notice that non-rigid deformation of human geometry typically occurs around a surface instead of in a volume, that is, nearby surface points on a parametric human model tend to have similar motion behavior. Thus we propose a novel motion parameterization technique that models the 3D human deformation in a 2D domain, enabling modeling of the motion field using 3D voxelized representation to accelerate the learning. This idea is similar to the displacement map and bump map [14, 15] in traditional computer graphics to represent detailed deformation on a 2D texture domain. We extend the technique of displacement map [14, 15] to represent human motions by restricting the originally 3D deformation field [40, 56, 59] on the 2D surface of a parametric human model, such as SMPL [43]. This technique allows us to use hybrid representations [48, 98] to model non-rigid deformation by reducing the memory footprint, largely boosting the convergence of the field.

Experiments demonstrate that our method significantly accelerates the optimization of neural human representations while being competitive with state-of-the-art human modeling methods on rendering quality. As shown in Figure 1, our model can be trained within 5 minutes to produce a volumetric video of a dynamic human from a 100-frame monocular video of a 512×512 resolution on an RTX 3090 GPU.

To summarize, our key contributions are:

- A novel part-based voxelized human representation for more efficient human body modeling.
- A 2D motion parameterization scheme for more efficient deformation field modeling.
- 100x speedup in optimization compared to previous neural human representations while maintaining competitive rendering quality.

2. Related work

Implicit neural representation and rendering. There have been many 3D scene representations, such as multi-view images [60, 78], textured meshes [38, 86], point clouds [1, 61], and voxels [42, 76]. Recently, some methods [11, 41, 44, 51, 77, 97, 104] propose implicit neural representations to represent scenes, which uses MLP networks to predict scene properties for any point in 3D space, such as occupancy [44, 67], signed distance [41, 51], and semantics [23, 104]. This enables them to describe continuous and high-resolution 3D scenes. To perform novel view synthesis, neural radiance field (NeRF) [47] models scenes as implicit fields of density and color. NeRF is optimized from images with volume rendering techniques, which produces impressive image synthesis results. Many works improve NeRF in various aspects, including rendering quality [3, 4], rendering speed [22, 27, 39, 63, 99], scene scale [64, 82, 89, 94], and reconstruction quality [50, 90, 96]. Some methods [18, 37, 52, 59] extend NeRF to dynamic scenes.

Human modeling. Reconstructing high-quality 3D human models is essential for synthesizing free-viewpoint videos of human performers. Traditional methods leverage multi-view stereo techniques [24, 71, 72] or depth fusion [13, 17, 80] to reconstruct human geometries, which require complicated hardware, such as dense camera arrays or depth sensors. To reduce the requirement of the capture equipment, some methods [2, 67, 68] train networks to learn human priors from datasets containing a large amount of 3D ground-truth human models, enabling them to infer human geometry and texture from even a single image. However, due to the limited diversity of training data, these methods do not generalize well to humans under complex poses. Recently, some methods [9, 46, 69, 87] model the shapes of dynamic humans as implicit neural representations and attempt to optimize them from human scans. Another line of works [28, 30, 34, 36, 40, 56, 58, 65, 92, 93, 95, 101–103, 105] exploits dynamic implicit neural representations and differentiable renderers to reconstruct 3D human models from

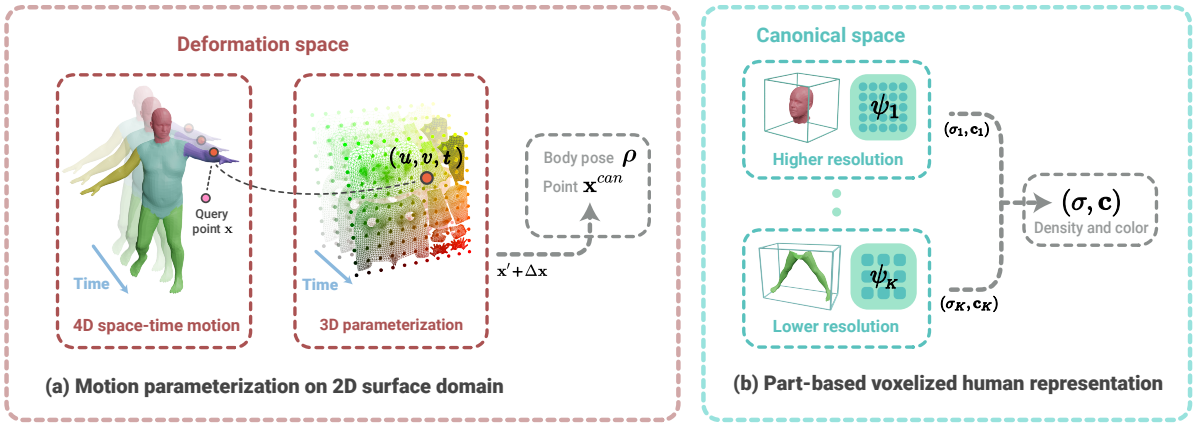

Figure 2. **Overview of the proposed representation.** Given a query point $\mathbf{x}$ at frame t, we find its nearest surface point on each human part of the SMPL mesh, which gives the blend weight $\mathbf{w}_k$ and the UV coordinates (u_k, v_k). Consider the k-th part. The motion field consists of an inverse LBS module and a residual deformation module. (a) The inverse LBS module takes body pose $\boldsymbol{\rho}$, blend weight $\mathbf{w}_k$, and query point $\mathbf{x}$ as input and outputs the transformed point $\mathbf{x}'$. The residual deformation module applies the multiresolution hash encoding (MHE) to (u_k, v_k, t) and uses an MLP network to regress the residual translation $\Delta\mathbf{x}$, which is added to $\mathbf{x}'$ to obtain the canonical point $\mathbf{x}^{\text{can}}$. (b) We then feed $\mathbf{x}^{\text{can}}$ to networks of k-th human part to predict the density σ_k and color $\mathbf{c}_k$. With $\{(\sigma_k, \mathbf{c}_k)\}_{k=1}^{K}$, we select the one with the largest density as the density and color of the query point.

videos. To represent dynamic humans, Neural Actor [40] augments the neural radiance field with the linear blend skinning model [35]. It additionally adopts a residual deformation field to better predict human motions. To overcome the inaccuracy of input human poses, [79, 93] optimize the parameters of human poses jointly with the human representations during training. These methods typically require a lengthy training process to produce high-quality human models. In contrast, we introduce a part-based voxelized human representation to model the canonical human body, which significantly accelerates the optimization process. Although [16, 46] have proposed part-based implicit functions, they focus on human shape modeling and do not show that the part-based representation can be used to reduce the training time.

Accelerating the optimization of neural representations. Many differentiable rendering-based methods [3, 39, 47] optimize a separate neural representation for each scene. The optimization process generally takes several hours on a modern GPU, which is time-consuming and costly to scale. Inspired by multi-view stereo matching [72], some methods [8, 12, 45, 73, 88, 91, 100] train a network on multi-view datasets to learn to infer radiance fields from input images. This enables them to quickly fine-tune neural representations to unseen scenes. [19, 62] leverage the auto-decoder [51] to capture the scene priors for efficient fine-tuning. [5, 83] utilize meta-learning techniques [20, 49] to initialize network parameters, thereby improving the training speed. Some methods [7, 10, 55, 70, 81, 98] attempt to design scene representations that support efficient training. [21, 48, 75, 84] augments the approximation ability of networks by designing encoding techniques. Multiresolution hash encoding [48] defines multiresolution feature vector arrays for a scene and uses the hash technique [85] to assign each input coordinate a feature vector as the encoded input, which significantly improves the training speed.

3. Methods

This paper aims to quickly create a volumetric video from a sparse multi-view video that captures a dynamic human. We assume that the cameras are calibrated and the human pose and foreground human mask of each image are provided. Section 3.1 introduces our dynamic human model that is comprised of a part-based voxelized human representation and a dimensionality reduction motion parameterization scheme. Section 3.2 discusses how to efficiently optimize the proposed representation. Finally, Section 3.3 provides implementation details.

3.1. Proposed human representation

As shown in Figure 2, our dynamic human representation consists of a motion parameterization field and a part-based voxelized human model. (a) For a query point $\mathbf{x}$, the motion parameterization field first transforms it to the canonical space correspondence $\mathbf{x}^{\text{can}}$ using the inverse LBS [35] algorithm and by parameterizing the 3D points onto 2D UV coordinates to predict the residual deformation $\Delta\mathbf{x}$. (b) Then, $\mathbf{x}^{\text{can}}$ is fed into the part-based voxelized human model in canonical space to predict and aggregate the density and color $(\sigma, \mathbf{c})$, where the canonical human body is decomposed into K parts, each of which is represented using an

MHE-augmented [48] NeRF network.

Motion parameterization on 2D surface domain. To regress the canonical correspondence $\mathbf{x}^{\text{can}}$ of a query point $\mathbf{x}$, we first find its nearest surface point $\mathbf{p}$ on the posed SMPL mesh. Using the strategy in [40], the blend weight $\mathbf{w}$ and UV coordinate (u, v) of surface point $\mathbf{p}$ are obtained from the SMPL model.

Given the blend weight $\mathbf{w}$ and UV coordinate (u, v), the motion field maps the query point $\mathbf{x}$ to the canonical space correspondence $\mathbf{x}^{\text{can}}$. The motion field is comprised of an inverse LBS module [35] and a residual deformation module. Given a query point $\mathbf{x}$ and blend weight $\mathbf{w}$, we use the inverse LBS module to transform it to the unposed space, which is defined as:

$$\Phi_{\text{LBS}}(\mathbf{x}, \mathbf{w}, \boldsymbol{\rho}) = \left(\sum_{j=1}^{J} w_{k,j} G_j \right)^{-1} \mathbf{x}, \tag{1}$$

where $\boldsymbol{\rho}$ denotes the human pose and $\{G_j\}_{j=1}^{J}$ are transformation matrices derived from $\boldsymbol{\rho}$ [43]. The detailed derivation of the inverse LBS algorithm can be found in the supplementary material.

The transformed point $\Phi_{\text{LBS}}(\mathbf{x}, \mathbf{w}, \boldsymbol{\rho})$ is then deformed to the human surface using the residual deformation module. Specifically, the current time t is first concatenated with the UV coordinate (u, v) to serve as the parameterization of the query point $\mathbf{x}$ at frame t. This motion parameterization is inspired by the displacement map and bump map techniques [14, 15] in the traditional Computer Graphics pipelines. It essentially reduces the dimensionality of a 4D space-time sequence down to the 3D surface-time domain utilizing the human deformation prior. Then, we apply the multiresolution hash encoding [48] ψ_{res} to (u, v, t) and forward the encoded input through a network MLP_{res} to regress the residual $\Delta\Phi(u, v, t)$. The full human motion at frame t is defined as:

$$\Delta\Phi(u, v, t) = \text{MLP}_{\text{res}}(\psi_{\text{res}}(u, v, t)), \tag{2}$$

$$\mathbf{x}^{\text{can}}(\mathbf{x}, \mathbf{w}, u, v, \boldsymbol{\rho}, t) = \Phi_{\text{LBS}}(\mathbf{x}, \mathbf{w}, \boldsymbol{\rho}) + \Delta\Phi(u, v, t). \tag{3}$$

There are two main observations that inspired us to use the (u, v, t) motion parameterization. First, we observe that a typical human motion happens at a surface level instead of a volumetric level. Near-surface points sharing similar UV coordinate of the parametric model shows similar motions. Utilizing this prior with a surface parameterization [14, 15], we can reduce the required 4D volumetric motion to the 3D surface-time domain, greatly decreasing the amount of information the deformation network has to learn. Based on a similar idea, [6] diffuses the surface motion to the full 3D space. Second, a naive (x, y, z, t) encoding would introduce quartic memory overhead on an explicitly defined voxel structure, which is intractable to use in practice. Instead, by parameterizing the motion to (u, v, t), we can reduce the memory footprint to a more practical cubic level. Experiments demonstrate that the motion parameterization scheme effectively reduces the dimensionality of the deformation field, thus greatly increasing the convergence rate of the human model.

Part-based voxelized human representation. Muller et al. [48] propose the multi-resolution hash encoding (MHE) to improve the approximation ability and training speed of implicit neural representations. MHE is defined on an explicit set of voxel grids of different resolutions. Given an input coordinate, it applies the hash encoding on each level and queries the corresponding voxel grid to trilinearly interpolate the feature of the input point for this level. Then, the concatenated multi-resolution feature is fed into a small MLP network to predict the target value. Note that [48] concatenates the features of multi-resolution hash encoding for the same point to mitigate the effect of hash collision, while our part-based voxelized human representation introduces spatially varying resolution to efficiently encode human parts with different complexity.

In contrast to [40, 56] which use a single neural radiance field (NeRF) to represent the canonical human model, we decompose the human body into multiple parts with different complexity and adopt a structured set of MHE-augmented NeRF with varying resolutions as the body representation. Specifically, we manually divide the human body into multiple parts based on a parametric human model (such as SMPL [43]), as shown in Figure 2. Note that other parametric human models [54,66] can also be used in our method. We use the blend weights defined in SMPL model [43] to decompose SMPL template mesh $\mathcal{M} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ represents the vertices and $\mathcal{E}$ represents the edges. Let the i-th vertice v_i have the blend weight w_i and for each part k we define Ω_k as the set of bones that belong to this part. The detailed setting of Ω_k can be found in the supplementary material. The mesh of the k-th part is defined as $\mathcal{M}_k = (\mathcal{V}_k, \mathcal{E}_k)$, where:

$$\mathcal{V}_k = \{v_i | \operatorname{argmax} w_i \in \Omega_k\}, \tag{4}$$

$$\mathcal{E}_k = \{(v_i, v_j) | v_i \in \mathcal{V}_k, v_j \in \mathcal{V}_k\}. \tag{5}$$

To regress the density and color of a query point $\mathbf{x}$, we first find the nearest surface point $\mathbf{p}_k$ on each human part $\mathcal{M}_k$ of the posed SMPL mesh. Using the strategy in [40], the blend weight $\mathbf{w}_k$ and UV coordinate (u_k, v_k) of surface point $\mathbf{p}_k$ are obtained from the SMPL model. With (u_k, v_k, t), we use the motion parameterization scheme defined in Section 3.1 to transform the query point to the space of the k-th human part. We predefine the parameters of the multiresolution hash encoding function ψ_k for the k-th part.

Given the transformed point, we first apply the multiresolution hash encoding to the transformed point and then feed the encoded point $\psi_k(\mathbf{x})$ to a small NeRF network to predict the density and color. The density network MLP_{σ_k} is defined as:

$$(\sigma_k, \mathbf{z}) = \text{MLP}_{\sigma_k}(\psi_k(\mathbf{x})), \tag{6}$$

where σ_k means the density and $\mathbf{z}$ is a feature vector. Then, we take the feature vector $\mathbf{z}$ and the viewing direction $\mathbf{d}$ as the input for the color regression. Similar to [52], a latent embedding $\boldsymbol{\ell}_t$ for each video frame t is introduced to model the temporally-varying appearance. The color network is defined as:

$$\mathbf{c}_k = \text{MLP}_{\mathbf{c}_k}(\mathbf{z}, \mathbf{d}, \boldsymbol{\ell}_t). \tag{7}$$

Finally, we have K predictions $\{(\sigma_k, \mathbf{c}_k)\}_{k=1}^{K}$. The density and color $(\sigma, \mathbf{c})$ of the query point $\mathbf{x}$ is calculated based on:

$$(\sigma, \mathbf{c}) = (\sigma_{k^*}, \mathbf{c}_{k^*}), \text{where } k^* = \underset{k}{\text{argmax}}\ \sigma_k. \tag{8}$$

In contrast to [40, 56] that represent the body with a single NeRF network, our part-based voxelized human representation can assign different densities of model parameters to different human parts with different complexity, thereby enabling us to efficiently distribute the representational power of the network. Experiments show that our proposed body representation significantly improves the rate of convergence.

3.2. Training

The proposed representation can be learned from sparse multi-view videos by minimizing the difference between rendered and observed images. The volume rendering technique [32, 47] is used to synthesize the pixel color. Give a pixel at frame t, we emit a camera ray and sample points along the ray. Then, the sampled points are fed into the dynamic human representation to predict their colors and densities, which are finally accumulated into the pixel color. During each training iteration, we randomly sample an image patch from the input image and compute the Mean Squared Error (MSE) loss and perceptual loss [31] to train the model parameters, which are defined as:

$$L_{\text{rgb}} = \|\tilde{I}_P - I_P\|_2 + \|F_{\text{vgg}}(\tilde{I}_P) - F_{\text{vgg}}(I_P)\|_2, \tag{9}$$

where $\tilde{I}_P$ is the rendered image patch, I_P is the ground truth image patch, and F_{vgg} extracts image features using the pretrained VGG network [31]. Ablation study demonstrates that perceptual loss is essential for rendering quality and fast training.

In addition to the image rendering loss, two regularization techniques are used to facilitate the learning of the neural representations. First, we apply the regularizer in [4] to concentrate densities on the human surface. Second, the residual deformation field is regularized to be small and smooth. More details of the loss terms are described in the supplementary material.

3.3. Implementation details

We adopt the Adam optimizer [33] with a learning rate of $5e^{-4}$. We train our model on an RTX 3090 GPU, which takes around 5 minutes to produce photorealistic results. Our method is implemented purely with the PyTorch framework [53] to demonstrate the effectiveness of our representation It also enables us to fairly compare with baseline methods [34, 56, 58] implemented in PyTorch. The details of the network architectures and hyper-parameters are presented in the supplementary material.

4. Experiments

4.1. Datasets

ZJU-MoCap [58] dataset is a widely-used benchmark for human modeling from videos. It provides foreground human masks and SMPL parameters. Following [93], we select 6 human subjects (377, 386, 387, 392, 393, 394) from this dataset to conduct our experiments. One camera is used for training, and the remaining cameras are used for evaluation. For each human subject, we select 1 frame every 5 frames and collect 100 frames for training. Please refer to the supplementary material for more detailed experiment settings of all characters.

MonoCap dataset contains four multi-view videos collected by [57] from the DeepCap dataset [26] and the DynaCap dataset [25]. It provides camera parameters and human masks. [57] additionally estimate the SMPL parameters for each image. We adopt the setting of training and test camera views in [57]. For each subject, 100 frames are selected for training, and we sample 1 frame every 5 frames. Detailed configurations of all sequences are described in the supplementary material.

4.2. Comparison with the state-of-the-art methods

Baselines. We compare our method with subject-specific optimization methods [56–58, 93], generalizable methods [34, 100]. All the baselines are implemented in pure PyTorch [53] for a fair comparison. Here we list only the average metric values of all selected characters on a dataset due to the size limit. We provide more detailed qualitative and quantitative comparisons in the supplementary material.

(1) Subject-specific optimization methods. Neural Body (NB) [58] anchors a set of latent codes to the SMPL mesh and regresses the radiance field from the posed latent codes. Animatable NeRF (AN) [56] deforms the canonical NeRF with the skeleton-driven framework and models non-rigid deformations by learning blend weight fields. [57] extend [56] with a signed distance field and pose-dependent deformation field to better model the residual deformation and geometric

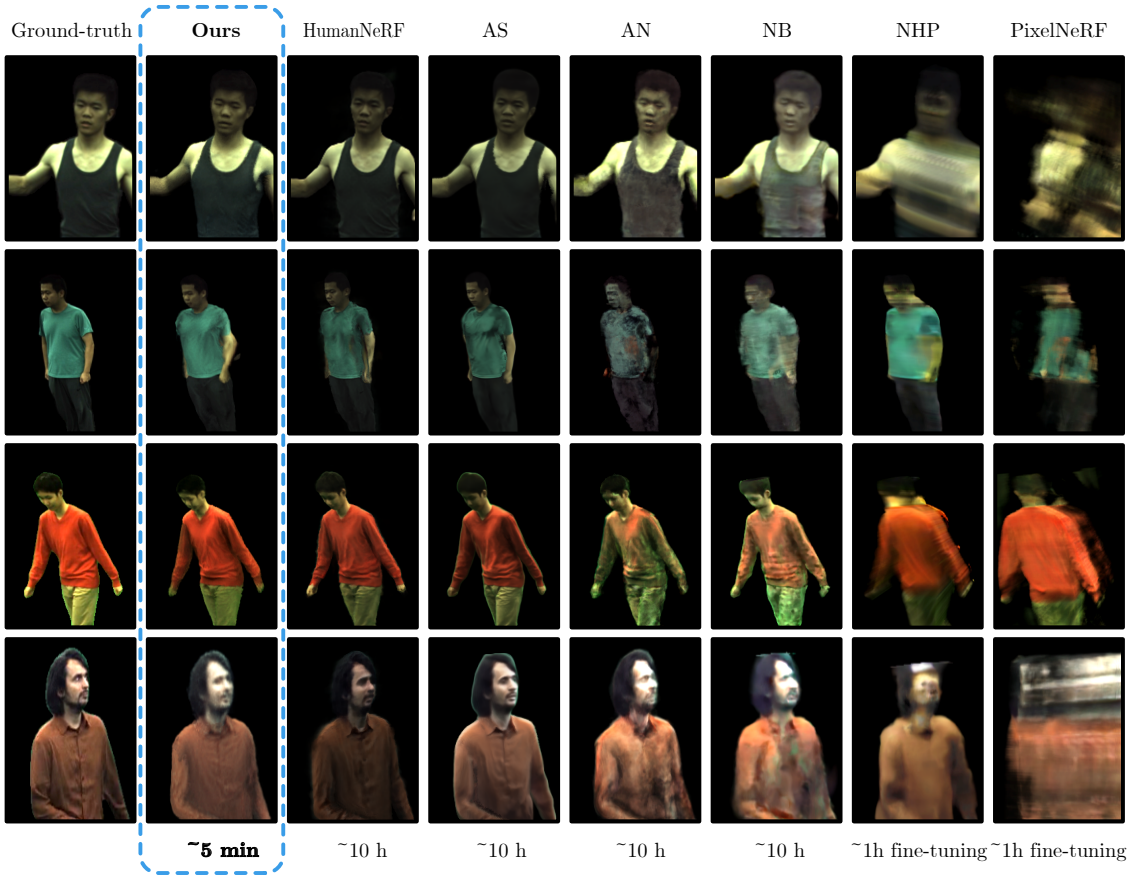

Figure 3. **Qualitative results of novel view synthesis on the ZJU-MoCap and MonoCap datasets.** Our method produces better rendering results while only requiring 1/100 of the training time. The bottom row lists the training time of each method.

details of dynamic humans. [93] optimizes for a volumetric representation of the person in a canonical space along with the estimated human pose.

(2) Generalizable methods. PixelNeRF [100] trains a network to infer the radiance field from a single image. Neural Human Performer (NHP) [34] anchors image features to vertices of the SMPL mesh and aggregates temporal features using a transformer, which are decoded into a human model. For each evaluated subject (e.g. one subject of MonoCap), we first pre-train the network on the other dataset (e.g. ZJU-MoCap) and then finetune it on the evaluated subject until it converges.

Results on the ZJU-MoCap dataset. Table 1 compares our method with NB [58], AN [56], PixelNeRF [100], NHP [34], HN [93] and AS [57] on novel view synthesis. Our proposed representation can be optimized within around 5 minutes to produce photorealistic rendering results, while [56–58, 93] require around 10 hours to finish training and [34, 100] require 10 hours of pretraining and 1 hour of fine-tuning. [57, 93] exhibit better results than [56, 58]. However, they all require a lengthy optimization process and fail to produce reasonable renderings in only 5 minutes because their models have not converged yet. Generalizable methods [34, 100] failed to render humans with reasonable shapes under the monocular setting. Our method achieves comparable results on all of the three evaluated metrics even when only trained in minutes, which shows the effectiveness of our novel human representation. We present qualitative results of our method and baselines in Figure 3.

Results on the MonoCap dataset. Table 1 summarizes the quantitative comparison between our method and other baselines on the MonoCap dataset. Our model again achieves competitive visual quality while only requiring 1/100 of the training time due to our efficient part-based voxelized human representation and effective motion parameterization scheme. Figure 3 indicates that our method can produce better appearance details than [56–58]. Although [56, 58] have shown impressive rendering results given 4-view videos, they do not perform well on monocular inputs. [58] implicitly aggregates the temporal information using structured latent

	Training Time	ZJU-MoCap			MonoCap		
		PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓
Ours	**~5 min**	**31.01**	0.971	38.45	32.61	**0.988**	16.68
HumanNeRF [93]	~10 h	30.66	0.969	**33.38**	**32.68**	0.987	15.52
AS [57]	~10 h	30.38	**0.975**	37.23	32.48	**0.988**	**13.18**
AN [56]	~10 h	29.77	0.965	46.89	31.07	0.985	19.47
NB [58]	~10 h	29.03	0.964	42.47	32.36	0.986	16.70
NHP [34]	~1 h fine-tuning	28.25	0.955	64.77	30.51	0.980	27.14
PixelNeRF [100]	~1 h fine-tuning	24.71	0.892	121.86	26.43	0.960	43.98

Table 1. **Quantitative comparison** of our method and baseline methods on the ZJU-MoCap and MonoCap datasets. We use **bold text** for the best and underlined text for the second best metric value across methods. Our method achieves the fastest training speed and shows competitive rendering results. Note that the NHP [34] and PixelNeRF [100] are additionally pretrained for 10 hours. LPIPS* = LPIPS $\times 10^3$.

(a) Ablation studies on proposed components.

	PSNR	SSIM	LPIPS*
Ours	**32.09**	**0.982**	**23.47**
Ours w/o Part	30.11	0.974	45.84
Ours w/o UV	31.40	0.979	30.99
Ours w/o Perc	30.55	0.976	44.33

(b) Ablation studies on variants of MLP_{res} input.

	PSNR	SSIM	LPIPS*
Ours	**32.09**	**0.982**	**23.47**
PE	31.94	0.981	26.76
XYZ-Code	31.32	0.979	31.19
XYZ-Pose	31.51	0.979	34.42

(c) Ablation studies on the part parameters.

	PSNR	SSIM	LPIPS*
Ours	**32.09**	**0.982**	**23.47**
Table size 2^{15}	30.69	0.976	35.67
Table size 2^{20}	31.18	0.978	33.58

Table 2. **Quantitative results of ablation studies** on the 377 sequence of ZJU-MoCap dataset. The description of each model can be found in Section 4.3. All models are trained for 5 minutes.

codes, which may not work well on monocular videos with complex human motions. [56] uses a learnable blend weight field to model human motion, which has a higher dimension and could be hard to converge well given single-view supervision. [57, 93] demonstrate similar visual quality and show that representing the human motion with the LBS model and residual deformation works particularly well, but their models require 100x more time to optimize.

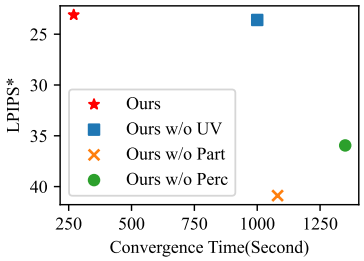

Figure 4. **Comparison of convergence LPIPS* and time needed for convergence among different variants of the proposed pipeline.** The description of each model can be found in Section 4.3. The proposed components accelerate the training and enhance the rendering quality significantly. LPIPS*=LPIPS $\times 10^3$.

4.3. Ablation Studies

We perform ablation studies on the "377" sequence of the ZJU-MoCap dataset to analyze how the proposed components affect the performance and training speed of our method.

4.3.1 Ablation Studies on Proposed Components.

Table 2(a) lists the quantitative result of ablation studies on our proposed components. All models are trained for 5 minutes. "Ours w/o Part" represents the canonical human body with a single MHE-augmented NeRF [48] network, which drastically degrades the rate of convergence. To keep the comparison fair, this variant of our method has a similar number of parameters (302M) to ours (286M). However, this change leads to a significant decrease in PSNR of 1.98dB because of its unwise design of considering all parts equally complex and wasting representational power. In "Ours w/o UV", the residual deformation network MLP_{res} takes hash encoded $(\mathbf{x}, t)$ as input, which observed a PSNR degradation from 32.09dB to 31.40dB with the same training time because of severe hash collision and limited resolution. "Ours w/o Perc" does not adopt the perceptual loss during training, which in turn increases the LPIPS distance. This comparison illustrates the importance of perceptual loss for visual fidelity. Figure 4 and Figure 5 provide a more intuitive qualitative comparison among variants of the proposed method.

Analysis of the part-based voxelized human representation. MHE [48] defines a multiresolution hash table of trainable features to embed input coordinates to a high-dimension space. We find that simply increasing the size

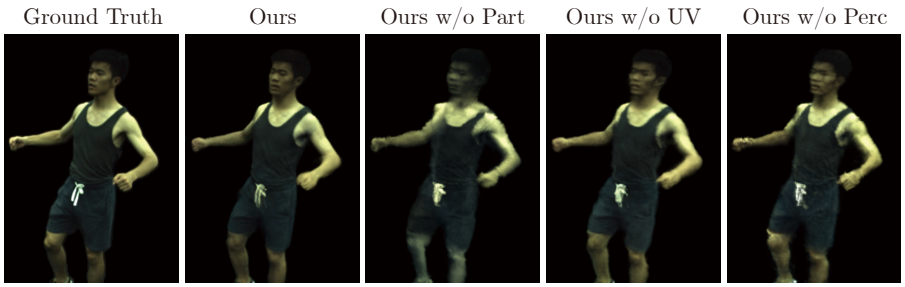

Figure 5. **Ablation studies on the 377 sequence of ZJU-MoCap dataset.** "Ours w/o part" means that we use an MHE-augmented NeRF network to represent the whole body. "Ours w/o UV" indicates that the residual deformation network MLP_{res} takes the hash encoded $(\mathbf{x}, t)$ as input. "Ours w/o Perc" represents that we do not use perceptual loss during training.

of the hash table does not always result in better performance with the same training time, because a bigger hash table leads to higher memory consumption and increases the time of each training iteration. The proposed part-based voxelized human representation allows us to adapt the hash table size according to the complexity of the human part, allowing us to efficiently represent the human body. Table 2(a) demonstrates the effectiveness of the part-based voxelized human representation. To further validate this representation, we design two variants that use the hash tables of size 2^{15} and 2^{20} in all human parts respectively. Table 2(c) summarizes the ablation studies, indicating that varying the model parameters in human parts improves the performance.

Analysis of the motion parameterization scheme. Table 2(a) shows that our model works better when the residual deformation network MLP_{res} takes parameterized 3D surface-time (u, v, t) coordinates as input, compared with taking 4D space-time $(\mathbf{x}, t)$ as input. Note that (u, v, t) makes MHE much more memory efficient than $(\mathbf{x}, t)$. To further validate the effectiveness of our motion parameterization, we additionally design three variants of the MLP_{res} input. (1) PE: positional encoded $(\mathbf{x}, t)$. (2) XYZ-Code: hash encoded $\mathbf{x}$ and a per-frame learnable latent code [57]. (3) XYZ-Pose: hash encoded $\mathbf{x}$ and the pose parameter [40, 93]. When taking the positional encoded $(\mathbf{x}, t)$ as input, we use a larger network for MLP_{res}. The results in Table 2(b) indicate that hash encoded (u, v, t) achieves the best performance.

Analysis of Robustness. To evaluate the robustness of the proposed system, we measure the time needed to achieve an evaluation PSNR of 30 for five times on the "377" sequence. This results in a training time with a mean value of $76.00s$ and a standard derivation of $13.56s$, showing the stability of the proposed method.

5. Limitations

Although our method can quickly reconstruct high-quality human models from videos, it has some limitations to be further addressed. First, our method currently relies on accurate SMPL parameters, which could be difficult to obtain under in-the-wild settings. It is interesting to utilize the techniques in [79, 93] to optimize the human pose parameters along with the training of human avatars. Second, we can only reconstruct foreground dynamic humans, while dynamic scenes typically include foreground and background entities. It might be plausible to combine our method with ST-NeRF [29] or MultiNB [74] to quickly reconstruct dynamic scenes containing foreground and background objects.

6. Conclusion

We introduce a novel dynamic human representation that can be quickly optimized from videos and used for generating free-viewpoint videos of the human performer. This representation consists of a part-based voxelized human model in the canonical space and a motion parameterization scheme that transforms points from the world space to the canonical space. The part-based voxelized human model decomposes the human body into multiple parts and represents each part with an MHE-augmented NeRF network, which efficiently distributes network representational power and significantly improves the training speed. When predicting the motion of a query point, the motion field reparameterizes the point coordinate as 2D surface-level UV coordinate, which effectively reduces the dimensionality of motion the network is required to model, resulting in a boost in convergence rate. Experiments demonstrate that our proposed representation can be optimized at 1/100 of the time of previous methods while still maintaining competitive rendering quality. We show that given a 100-frame monocular video of a 512×512 resolution, our method can produce photorealistic free-viewpoint videos in minutes on an RTX 3090 GPU.

Acknowledgements. This work was supported by the Key Research Project of Zhejiang Lab (No. K2022PG1BB01), NSFC (No. 62172364), and Information Technology Center and State Key Lab of CAD&CG, Zhejiang University.

References

[1] Kara-Ali Aliev, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov, and Victor Lempitsky. Neural point-based graphics. In *ECCV*, 2020. 2

[2] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. *arXiv preprint arXiv:2204.08906*, 2022. 2

[3] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 2, 3

[4] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *arXiv preprint arXiv:2111.12077*, 2021. 2, 5

[5] Alexander Bergman, Petr Kellnhofer, and Gordon Wetzstein. Fast training of neural lumigraph representations using meta learning. *Advances in Neural Information Processing Systems*, 34, 2021. 3

[6] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration. In *NeurIPS*, 2020. 4

[7] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. *arXiv preprint arXiv:2203.09517*, 2022. 3

[8] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021. 3

[9] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11594–11604, 2021. 2

[10] Yue Chen, Xuan Wang, Qi Zhang, Xiaoyu Li, Xingyu Chen, Yu Guo, Jue Wang, and Fei Wang. Uv volumes for real-time rendering of editable free-view human performance. *arXiv preprint arXiv:2203.14402*, 2022. 3

[11] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *CVPR*, 2019. 2

[12] Julian Chibane, Aayush Bansal, Verica Lazova, and Gerard Pons-Moll. Stereo radiance fields (srf): Learning view synthesis for sparse views of novel scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7911–7920, 2021. 3

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

[14] Robert L Cook. Shade trees. In *Proceedings of the 11th annual conference on Computer graphics and interactive techniques*, pages 223–231, 1984. 2, 4

[15] Robert L Cook, Loren Carpenter, and Edwin Catmull. The reyes image rendering architecture. *ACM SIGGRAPH Computer Graphics*, 21(4):95–102, 1987. 2, 4

[16] Boyang Deng, John P Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Nasa neural articulated shape approximation. In *European Conference on Computer Vision*, pages 612–628. Springer, 2020. 3

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

[18] Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B Tenenbaum, and Jiajun Wu. Neural radiance flow for 4d view synthesis and video processing. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14304–14314. IEEE Computer Society, 2021. 2

[19] Emilien Dupont, Hyunjik Kim, SM Eslami, Danilo Rezende, and Dan Rosenbaum. From data to functa: Your data point is a function and you should treat it like one. *arXiv preprint arXiv:2201.12204*, 2022. 3

[20] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017. 3

[21] Xuan Gao, Chenglai Zhong, Jun Xiang, Yang Hong, Yudong Guo, and Juyong Zhang. Reconstructing personalized semantic facial nerf models from monocular video. *arXiv preprint arXiv:2210.06108*, 2022. 3

[22] Stephan J Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien Valentin. Fastnerf: High-fidelity neural rendering at 200fps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14346–14355, 2021. 2

[23] Haoyu Guo, Sida Peng, Haotong Lin, Qianqian Wang, Guofeng Zhang, Hujun Bao, and Xiaowei Zhou. Neural 3d scene reconstruction with the manhattan-world assumption. In *CVPR*, 2022. 2

[24] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM TOG*, 2019. 2

[25] Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Real-time deep dynamic characters. In *SIGGRAPH*, 2021. 5

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

[27] Peter Hedman, Pratul P Srinivasan, Ben Mildenhall, Jonathan T Barron, and Paul Debevec. Baking neural radiance fields for real-time view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5875–5884, 2021. 2

[28] Tao Hu, Tao Yu, Zerong Zheng, He Zhang, Yebin Liu, and Matthias Zwicker. Hvtr: Hybrid volumetric-textural rendering for human avatars. *arXiv preprint arXiv:2112.10203*, 2021. 2

[29] Zhang Jiakai, Liu Xinhang, Ye Xinyi, Zhao Fuqiang, Zhang Yanshun, Wu Minye, Zhang Yingliang, Xu Lan, and Yu Jingyi. Editable free-viewpoint video using a layered neural representation. In *SIGGRAPH*, 2021. 8

[30] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. *arXiv preprint arXiv:2201.12792*, 2022. 2

[31] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 5

[32] James T Kajiya and Brian P Von Herzen. Ray tracing volume densities. *ACM SIGGRAPH computer graphics*, 18(3):165–174, 1984. 5

[33] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[34] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 5, 6, 7

[35] John P Lewis, Matt Cordner, and Nickson Fong. Pose space deformation: a unified approach to shape interpolation and skeleton-driven deformation. In *SIGGRAPH*, 2000. 2, 3, 4

[36] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhofer, Jurgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. *arXiv preprint arXiv:2206.08929*, 2022. 2

[37] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6508, 2021. 2

[38] Yiyi Liao, Katja Schwarz, Lars Mescheder, and Andreas Geiger. Towards unsupervised learning of generative models for 3d controllable image synthesis. In *CVPR*, 2020. 2

[39] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *Advances in Neural Information Processing Systems*, 33:15651–15663, 2020. 2, 3

[40] 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. In *SIGGRAPH Asia*, 2021. 2, 3, 4, 5, 8

[41] Shaohui Liu, Yinda Zhang, Songyou Peng, Boxin Shi, Marc Pollefeys, and Zhaopeng Cui. Dist: Rendering deep implicit signed distance function with differentiable sphere tracing. In *CVPR*, 2020. 2

[42] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. In *SIGGRAPH*, 2019. 2

[43] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM TOG*, 2015. 2, 4

[44] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, 2019. 2

[45] Marko Mihajlovic, Aayush Bansal, Michael Zollhoefer, Siyu Tang, and Shunsuke Saito. Keypointnerf: Generalizing image-based volumetric avatars using relative spatial encoding of keypoints. In *European Conference on Computer Vision*, pages 179–197. Springer, 2022. 3

[46] Marko Mihajlovic, Shunsuke Saito, Aayush Bansal, Michael Zollhoefer, and Siyu Tang. COAP: Compositional articulated occupancy of people. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3

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

[48] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. In *SIGGRAPH*, 2022. 2, 3, 4, 7

[49] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018. 3

[50] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5589–5599, 2021. 2

[51] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019. 2, 3

[52] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 2, 5

[53] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 5

[54] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 4

[55] Bo Peng, Jun Hu, Jingtao Zhou, and Juyong Zhang. Selfnerf: Fast training nerf for human from monocular self-rotating video. *arXiv preprint arXiv:2210.01651*, 2022. 3

[56] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14314–14323, 2021. 2, 4, 5, 6, 7

[57] Sida Peng, Shangzhan Zhang, Zhen Xu, Chen Geng, Boyi Jiang, Hujun Bao, and Xiaowei Zhou. Animatable neural implicit surfaces for creating avatars from videos. *arXiv preprint arXiv:2203.08133*, 2022. 5, 6, 7, 8

[58] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021. 1, 2, 5, 6, 7

[59] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 2

[60] Charles R Qi, Hao Su, Matthias Nießner, Angela Dai, Mengyuan Yan, and Leonidas J Guibas. Volumetric and multi-view cnns for object classification on 3d data. In *CVPR*, 2016. 2

[61] Amit Raj, Julian Tanke, James Hays, Minh Vo, Carsten Stoll, and Christoph Lassner. Anr: Articulated neural rendering for virtual avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3722–3731, 2021. 2

[62] Eduard Ramon, Gil Triginer, Janna Escur, Albert Pumarola, Jaime Garcia, Xavier Giro-i Nieto, and Francesc Moreno-Noguer. H3d-net: Few-shot high-fidelity 3d head reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5620–5629, 2021. 3

[63] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14335–14345, 2021. 2

[64] Konstantinos Rematas, Andrew Liu, Pratul P Srinivasan, Jonathan T Barron, Andrea Tagliasacchi, Thomas Funkhouser, and Vittorio Ferrari. Urban radiance fields. *arXiv preprint arXiv:2111.14643*, 2021. 2

[65] Edoardo Remelli, Timur Bagautdinov, Shunsuke Saito, Chenglei Wu, Tomas Simon, Shih-En Wei, Kaiwen Guo, Zhe Cao, Fabian Prada, Jason Saragih, et al. Drivable volumetric avatars using texel-aligned features. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–9, 2022. 2

[66] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM ToG*, 2017. 4

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

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

[69] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J Black. Scanimate: Weakly supervised learning of skinned clothed avatar networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2886–2897, 2021. 2

[70] Vishwanath Saragadam, Jasper Tan, Guha Balakrishnan, Richard G Baraniuk, and Ashok Veeraraghavan. Miner: Multiscale implicit neural representations. *arXiv preprint arXiv:2202.03532*, 2022. 3

[71] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 2

[72] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016. 2, 3

[73] Ruizhi Shao, Hongwen Zhang, He Zhang, Yanpei Cao, Tao Yu, and Yebin Liu. Doublefield: Bridging the neural surface and radiance fields for high-fidelity human rendering. *arXiv preprint arXiv:2106.03798*, 2021. 3

[74] Qing Shuai, Chen Geng, Qi Fang, Sida Peng, Wenhao Shen, Xiaowei Zhou, and Hujun Bao. Novel view synthesis of human interactions from sparse multi-view videos. In *SIGGRAPH Conference Proceedings*, 2022. 8

[75] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020. 3

[76] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *CVPR*, 2019. 2

[77] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *NeurIPS*, 2019. 2

[78] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 945–953, 2015. 2

[79] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. *Advances in Neural Information Processing Systems*, 34, 2021. 3, 8

[80] Zhuo Su, Lan Xu, Zerong Zheng, Tao Yu, Yebin Liu, et al. Robustfusion: Human volumetric capture with data-driven visual cues using a rgbd camera. In *ECCV*, 2020. 2

[81] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. *arXiv preprint arXiv:2111.11215*, 2021. 3

[82] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. *arXiv preprint arXiv:2202.05263*, 2022. 2

[83] Matthew Tancik, Ben Mildenhall, Terrance Wang, Divi Schmidt, Pratul P Srinivasan, Jonathan T Barron, and Ren Ng. Learned initializations for optimizing coordinate-based neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2846–2855, 2021. 3

[84] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020. 3

[85] Matthias Teschner, Bruno Heidelberger, Matthias Müller, Danat Pomerantes, and Markus H Gross. Optimized spatial hashing for collision detection of deformable objects. In *Vmv*, volume 3, pages 47–54, 2003. 3

[86] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM TOG*, 2019. 2

[87] Garvita Tiwari, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Neural-gif: Neural generalized implicit functions for animating people in clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11708–11718, 2021. 2

[88] Alex Trevithick and Bo Yang. Grf: Learning a general radiance field for 3d representation and rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15182–15192, 2021. 3

[89] Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs. *arXiv preprint arXiv:2112.10703*, 2021. 2

[90] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *NeurIPS*, 2021. 2

[91] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021. 3

[92] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. Arah: animatable volume rendering of articulated human sdfs. *arXiv preprint arXiv:2210.10036*, 2022. 2

[93] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. *arXiv preprint arXiv:2201.04127*, 2022. 1, 2, 3, 5, 6, 7, 8

[94] Yuanbo Xiangli, Linning Xu, Xingang Pan, Nanxuan Zhao, Anyi Rao, Christian Theobalt, Bo Dai, and Dahua Lin. Citynerf: Building nerf at city scale. *arXiv preprint arXiv:2112.05504*, 2021. 2

[95] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. *Advances in Neural Information Processing Systems*, 34, 2021. 2

[96] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *NeurIPS*, 2021. 2

[97] Jianglong Ye, Yuntao Chen, Naiyan Wang, and Xiaolong Wang. Gifs: Neural implicit function for general shape representation. *arXiv preprint arXiv:2204.07126*, 2022. 2

[98] Alex Yu, Sara Fridovich-Keil, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. *arXiv preprint arXiv:2112.05131*, 2021. 2, 3

[99] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5752–5761, 2021. 2

[100] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 2, 3, 5, 6, 7

[101] Ruiqi Zhang and Jie Chen. Ndf: Neural deformable fields for dynamic human modelling. *arXiv preprint arXiv:2207.09193*, 2022. 2

[102] Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. Humannerf: Generalizable neural human radiance field from sparse inputs. *arXiv preprint arXiv:2112.02789*, 2021. 2

[103] Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. Structured local radiance fields for human avatar modeling. *arXiv preprint arXiv:2203.14478*, 2022. 2

[104] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *ICCV*, 2021. 2

[105] Yihao Zhi, Shenhan Qian, Xinhao Yan, and Shenghua Gao. Dual-space nerf: Learning animatable avatars and scene lighting in separate spaces. *arXiv preprint arXiv:2208.14851*, 2022. 2