InstantAvatar: Learning Avatars from Monocular Video in 60 Seconds

Tianjian Jiang[1]*, Xu Chen[1,2]*, Jie Song[†1], Otmar Hilliges[1]
[1] ETH Zürich [2] Max Planck Institute for Intelligent Systems, Tübingen

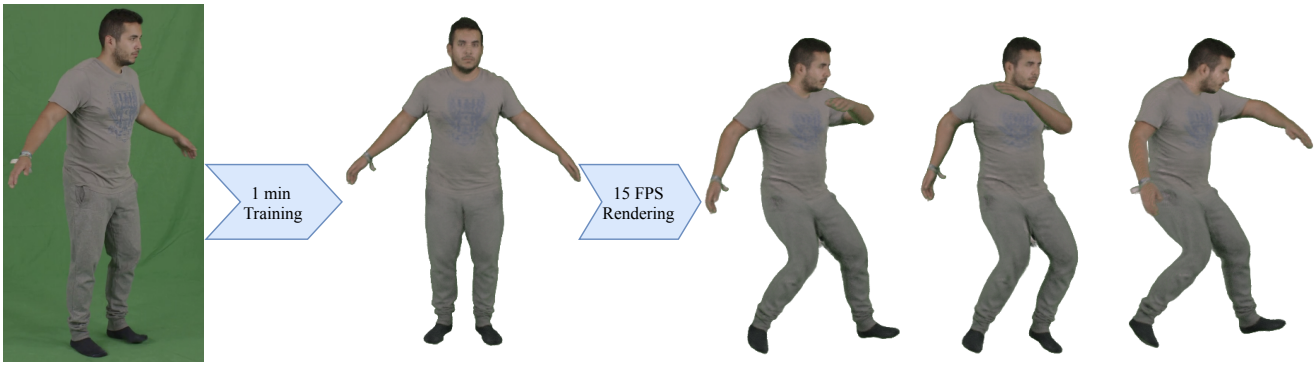

Figure 1. **InstantAvatar:** we propose a system that can reconstruct animatable high-fidelity human avatars from a monocular video within 60 seconds, providing poses and masks, and can animate and render the model at 15 FPS at 540×540 resolution. To achieve this we integrate accelerated neural radiance fields, originally designed for rigid scenes, with a fast correspondence search module for articulation. An efficient empty-space skipping strategy further speeds up training and inference, enabling near-instant avatar learning.

Abstract

In this paper, we take one step further towards real-world applicability of monocular neural avatar reconstruction by contributing InstantAvatar, a system that can reconstruct human avatars from a monocular video within seconds, and these avatars can be animated and rendered at an interactive rate. To achieve this efficiency we propose a carefully designed and engineered system, that leverages emerging acceleration structures for neural fields, in combination with an efficient empty-space skipping strategy for dynamic scenes. We also contribute an efficient implementation that we will make available for research purposes. Compared to existing methods, InstantAvatar converges 130× faster and can be trained in minutes instead of hours. It achieves comparable or even better reconstruction quality and novel pose synthesis results. When given the same time budget, our method significantly outperforms SoTA methods. InstantAvatar can yield acceptable visual quality in as little as 10 seconds training time. For code and more demo results, please refer to https://ait.ethz.ch/InstantAvatar.

*Equal contribution.
†Corresponding author

1. Introduction

Creating high-fidelity digital humans is important for many applications including immersive telepresence, AR/VR, 3D graphics, and the emerging metaverse. Currently acquiring personalized avatars is an involved process that typically requires the use of calibrated multi-camera systems and incurs significant computational cost. In this paper, we embark on the quest to build a system for the learning of 3D virtual humans from monocular video alone that is lightweight enough to be widely deployable and fast enough to allow for walk-up and use scenarios.

The emergence of powerful neural fields has enabled a number of methods for the reconstruction of animatable avatars from monocular videos of moving humans [1, 2, 6, 49, 62]. These methods typically model human shape and appearance in a pose-independent canonical space.

To reconstruct the model from images that depict humans in different poses, such methods must use animation (e.g. skinning) and rendering algorithms, to deform and render the model into posed space in a differentiable way. This mapping between posed and canonical space allows optimization of network weights by minimizing the difference between the generated pixel values and real images. Especially methods that leverage neural radiance

fields (NeRFs) [40] as the canonical model has demonstrated high-fidelity avatar reconstruction results. However, due to the dual need for differentiable deformation modules and for volume rendering, these models require hours of training time and cannot be rendered at interactive rates, prohibiting their broader application.

In this paper, we aim to take a further step toward real-world applicability of monocular neural avatar reconstruction by contributing a method that takes no longer for reconstruction, than it takes to capture the input video. To this end, we propose InstantAvatar, a system that reconstructs high-fidelity avatars within 60 seconds, instead of hours, given a monocular video, pose parameters and masks. Once learned the avatar can be animated and rendered at interactive rates. Achieving such a speed-up is clearly a challenging task that requires careful method design, requires fast differentiable algorithms for rendering and articulation, and requires efficient implementation.

Our simple yet highly efficient pipeline combines several key components. First, to learn the canonical shape and appearance we leverage a recently proposed neural radiance field variant [42]. Instant-NGP [42] accelerates neural volume rendering by replacing multi-layer perceptrons (MLPs) with a more efficient hash table as data structure. However, because the spatial features are represented explicitly, Instant-NGP is limited to rigid objects. Second, to enable learning from posed observations and to be able to animate the avatar, we interface the canonical NeRF with an efficient articulation module, Fast-SNARF [7], which efficiently derives a continuous deformation field to warp the canonical radiance field into the posed space. Fast-SNARF is orders of magnitude faster compared to its slower predecessor [9].

Finally, simply integrating existing acceleration techniques is not sufficient to yield the desired efficiency. With acceleration structures for the canonical space and a fast articulation module in place, rendering the actual volume becomes the computational bottleneck. To compute the color of a pixel, standard volume rendering needs to query and accumulate densities of hundreds of points along the ray. A common approach to accelerating this is to maintain an occupancy grid to skip samples in the empty space. However, such an approach assumes rigid scenes and can not be applied to dynamic scenes such as humans in motion.

We propose an empty space skipping scheme that is designed for dynamic scenes with known articulation patterns. At inference time, for each input body pose, we sample points on a regular grid in posed space and map them back to the canonical model to query densities. Thresholding these densities yields an occupancy grid in canonical space, which can then be used to skip empty space during volume rendering. For training, we maintain a shared occupancy grid over all training frames, recording the union of occupied regions over individual frames. This occupancy grid is updated every few training iterations with the densities of randomly sampled points, in the posed space of randomly sampled frames. This scheme balances computational efficiency and rendering quality.

We evaluate our method on both synthetic and real monocular videos of moving humans and compare it with state-of-the-art methods on monocular avatar reconstruction. Our method achieves on-par reconstruction quality and better animation quality in comparison to SoTA methods, while only requiring minutes of training time instead of more than 10 hours. When given the same time budget, our method significantly outperforms SoTA methods. We also provide an ablation study to demonstrate the effect of our system's components on speed and accuracy.

2. Related Work

3D Human Reconstruction Reconstructing 3D human appearance and shape is a long-standing problem. High-quality reconstruction has been achieved in [12, 15, 19, 36] by fusing observations from a dense array of cameras or depth sensors. The expensive hardware requirement limits such methods to professional settings. Recent work [1, 2, 17, 20, 21, 28, 65] demonstrates 3D human reconstruction from a monocular video by leveraging personalized or generic template mesh models such as SMPL [35]. These methods reconstruct 3D humans by deforming the template to fit 2D joints and silhouettes. However, personalized template mesh might not be available in many scenarios and generic template mesh cannot model high-fidelity details and different clothing typologies.

Recently, neural representations [37, 41, 45, 46] have emerged as a powerful tool to model 3D humans [3, 6, 8, 10, 11, 13, 14, 22–26, 30, 31, 34, 38, 39, 39, 43, 44, 48, 49, 52, 53, 57, 59–63, 63, 64, 67, 69, 70]. Using neural representations, many works [6, 18, 26, 27, 30, 34, 43, 48, 49, 61, 62, 64, 69] can directly reconstruct high fidelity neural human avatars from a sparse set of views or a monocular video without pre-scanning personalized template. These methods model 3D human shape and appearance via neural radiance field [40] or signed distance and texture field in a pose-independent canonical space and then deform and render the model into various body poses in order to learn from posed observations. While achieving impressive quality and can learn avatars from a monocular video, these methods suffer from slow training and rendering speed due to the slow speed of the canonical representation as well as deformation algorithms. Our method addresses this issue and enables learning avatars within minutes.

Accelerating Neural Radiance Field Several methods have been proposed to improve the training and inference speed of neural representations [5, 16, 29, 32, 33, 42, 51,

54–56, 66]. The core idea is to replace MLPs in neural representations with more efficient representations. A few works [33, 54, 66] propose to use voxel grids to represent neural fields and achieve fast training and inference speed. Instant-NGP [42] further replaces dense voxels with a multi-resolution hash table, which is more memory efficient and hence can record high-frequency details. Besides improving the efficiency of the representation, several works [29, 32, 42] also improve the rendering efficiency by skipping empty space via an occupancy grid to further increase training and inference speed.

While achieving impressive quality and training efficiency, these methods are specifically designed for rigid objects. Generalizing these methods to non-rigid objects is not straightforward. We combine Instant-NGP with a recent articulation algorithm to enable animation and learning from posed observations. In addition, we propose an empty space skinning scheme for dynamic articulated humans.

3. Method

Given a monocular video of a moving human, our primary goal is to reconstruct a 3D human avatar within a tight computational budget. In this section, we first describe the preliminaries that our method is based on (Sec. 3.1), which include an accelerated neural radiance field that we use to model the appearance and shape in canonical space and an efficient articulation module to deform the canonical radiance field into posed space. We then describe our implementation of the volumetric renderer to produce images from the radiance fields in an efficient manner (Sec. 3.2). To avoid inefficient sampling of empty space, we leverage the observation that the 3D bounding box around the human body is dominated by empty space. We then propose an empty space skipping scheme specifically designed for humans (Sec. 3.3). Finally, we discuss training objectives and regularization strategies (Sec. 3.4).

3.1. Preliminaries

Efficient Canonical Neural Radiance Field We model human shape and appearance in a canonical space using a radiance field $\mathbf{f}_{\sigma_f}$, which predicts the density σ and color c of each 3D point $\mathbf{x}$ in the canonical space:

$$\mathbf{f}_{\sigma_f} : \mathbb{R}^3 \rightarrow \mathbb{R}^+, \mathbb{R}^3 \tag{1}$$

$$\mathbf{x} \mapsto \sigma, c \tag{2}$$

where σ_f are the parameters of the radiance field.

We use Instant-NGP [42] to parameterize $\mathbf{f}_{\sigma_f}$, which achieves fast training and inference speed by using a hash table to store feature grids at different coarseness scales. To predict the texture and geometry properties of a query point in space, they read and tri-linearly interpolate the features at its neighboring grid points and then concatenate the interpolated features at different levels. The concatenated features are finally decoded with a shallow MLP.

Articulating Radiance Fields To create animations and to learn from posed images, we need to generate deformed radiance fields in target poses $\mathbf{f}'_{\sigma_f}$. The posed radiance field is defined as

$$\mathbf{f}'_{\sigma_f} : \mathbb{R}^3 \rightarrow \mathbb{R}^+, \mathbb{R}^3 \tag{3}$$

$$\mathbf{x}' \mapsto \sigma, c, \tag{4}$$

which outputs color and density for each point in posed space. We use a skinning weight field $\mathbf{w}$ in canonical space to model articulation, with σ_w being its parameters:

$$\mathbf{w}_{\sigma_w} : \mathbb{R}^3 \rightarrow \mathbb{R}^{n_b}, \tag{5}$$

$$\mathbf{x} \mapsto w_1, ..., w_{n_b}. \tag{6}$$

where n_b is the number of bones in the skeleton. To avoid the computational cost of [9], [7] represents this skinning weight field as a low-resolution voxel grid. The value of each grid point is determined as the skinning weights of its nearest vertex on the SMPL [35] model. With this the canonical skinning weight field and target bone transformations $\mathbf{B} = \{\mathbf{B}_1, ..., \mathbf{B}_{n_b}\}$, a point $\mathbf{x}$ in canonical space is transformed to deformed space $\mathbf{x}'$ via linear blend skinning as follows:

$$\mathbf{x}' = \textstyle\sum_{i=1}^{n_b} w_i \mathbf{B}_i \mathbf{x} \tag{7}$$

The canonical correspondences $\mathbf{x}^*$ of a deformed point $\mathbf{x}'$ are defined by the inverse mapping of Equation. 7. The key is to establish the mapping from points in posed space $\mathbf{x}'$ to their correspondences in the canonical space $\mathbf{x}^*$. This is efficiently derived by root-finding in Fast-SNARF [7]. The posed radiance field $\mathbf{f}'_{\sigma_f}$ can then be determined as $\mathbf{f}'_{\sigma_f}(\mathbf{x}') = \mathbf{f}_{\sigma_f}(\mathbf{x}^*)$.

3.2. Rendering Radiance Fields

The articulated radiance field $\mathbf{f}'_{\sigma_f}$ can be rendered into novel views via volume rendering. Given a pixel, we cast a ray $\mathbf{r} = \mathbf{o} + t\mathbf{d}$ with $\mathbf{o}$ being the camera center and $\mathbf{d}$ being the ray direction. We sample N points $\{\mathbf{x}'_i\}^N$ along the ray between the near and far bound, and query the color and density of each point from the articulated radiance field $\mathbf{f}'_{\sigma_f}$ by mapping $\{\mathbf{x}'_i\}^N$ back to the canonical space and querying from the canonical NeRF model $\mathbf{f}_{\sigma_f}$, as illustrate in Fig. 2. We then accumulate queried radiance and density along the ray to get the pixel color C

$$C = \sum_{i=1}^{N} \alpha_i \prod_{j<i} (1 - \alpha_j) c_i, \text{with } \alpha_i = 1 - \exp(\sigma_i \delta_i) \tag{8}$$

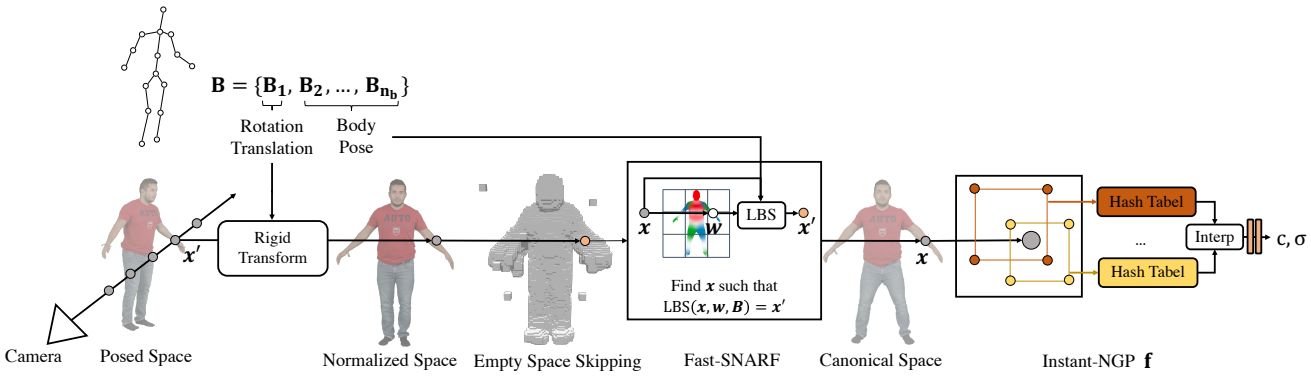

Figure 2. **Method Overview.** For each frame, we sample points along the rays in posed space. We then transform these points into a normalized space where the global orientation and translation are removed, we then filter points in empty space using our occupancy grid. The remaining points are deformed to canonical space using an articulation module and then fed into the canonical neural radiance field to evaluate the color and density.

where $\delta_i = \|\mathbf{x}'_{i+1} - \mathbf{x}'_i\|$ is the distance between samples.

While the acceleration modules of Sec. 3.1 already achieve significant speed-up over the vanilla variants (NeRF [40], SNARF [9]), the rendering itself now becomes the bottleneck. In this paper, we optimize the process of neural rendering, specifically for the use-case of dynamic humans.

3.3. Empty Space Skipping for Dynamic Objects

We note that the 3D bounding box surrounding the human body is dominated by empty space due to the articulated structure of 3D human limbs. This results in a large amount of redundant sample queries during rendering and hence significantly slows down rendering. For rigid objects, this problem is eliminated by caching a coarse occupancy grid and skipping samples within non-occupied grid cells. However, for dynamic objects, the exact location of empty space varies across different frames, depending on the pose.

Inference Stage At inference time, for each input body pose, we sample points on a $64 \times 64 \times 64$ grid in posed space and query their densities from the posed radiance field $\mathbf{f}'_{\sigma_f}$. We then threshold these densities into binary occupancy values. To remove cells that have been falsely labeled as empty, due to the low spatial resolution, we dilate the occupied region to fully cover the subject. Due to the low resolution of this grid and the large amount of queries required to render an image, the overhead to construct such an occupancy grid is negligible. During volumetric rendering, for point samples inside the non-occupied cells, we directly set their density to zero without querying the posed radiance field $\mathbf{f}'_{\sigma_f}$. This reduces unnecessary computation to a minimum and hence improves the inference speed.

Training Stage During training, however, the overhead to construct such an occupancy grid at each training iteration is no longer negligible. To avoid this overhead, we construct a *single* occupancy grid for the entire sequence by recording the union of occupied regions in each of the individual frames. Specifically, we build an occupancy grid at the start of training and update it every k iterations, by taking the moving average of the current occupancy values and the densities queried from the posed radiance field $\mathbf{f}'_{\sigma_f}$ at the current iteration. Note that this occupancy grid is defined in a normalized space where the global orientation and translation are factored out so that the union of the occupied space is as tight as possible and hence unnecessary queries are further reduced.

3.4. Training Losses

We train our model by minimizing the robust Huber loss ρ between the predicted color of the pixels C and the corresponding ground-truth color C_{gt}:

$$\mathcal{L}_{\text{rgb}} = \rho(\|C - C_{gt}\|) \tag{9}$$

In addition, we assume an estimate of the human mask is available and apply a loss on the rendered 2D alpha values, in order to reduce floating artifacts in space.

$$\mathcal{L}_{\text{alpha}} = \|\alpha - \alpha_{gt}\|_1 \tag{10}$$

Hard Surface Regularization Following [50], we add further regularization to encourage the NeRF model to predict solid surfaces:

$$\mathcal{L}_{\text{hard}} = -\log(\exp^{-|\alpha|} + \exp^{-|\alpha-1|}) + \text{const.} \tag{11}$$

where const. is a constant to ensure loss value to be non-negative. Encouraging solid surfaces helps to speed up rendering because we can terminate rays early once the accumulated opacity reaches 1.

	male-3-casual			male-4-casual			female-3-casual			female-4-casual		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Neural Body [49] (∼ 14 hours)	24.94	0.9428	0.0326	24.71	0.9469	0.0423	23.87	0.9504	0.0346	24.37	0.9451	0.0382
Anim-NeRF [6] (∼ 13 hours)	29.37	0.9703	**0.0168**	**28.37**	0.9605	**0.0268**	**28.91**	**0.9743**	**0.0215**	28.90	0.9678	**0.0174**
Ours (1 minute)	**29.65**	**0.9730**	0.0192	27.97	**0.9649**	0.0346	27.90	0.9722	0.0249	**28.92**	**0.9692**	0.0180
Anim-NeRF [6] (5 minutes)	23.17	0.9266	0.0784	22.30	0.9235	0.0911	22.37	0.9311	0.0784	23.18	0.9292	0.0687
Ours (5 minutes)	**29.53**	**0.9716**	**0.0155**	**27.67**	**0.9626**	**0.0307**	**27.66**	**0.9709**	**0.0210**	**29.11**	**0.9683**	**0.0167**
Anim-NeRF [6] (3 minutes)	19.75	0.8927	0.1286	20.66	0.8986	0.1414	19.77	0.9003	0.1255	20.20	0.9044	0.1109
Ours (3 minutes)	**29.58**	**0.9719**	**0.0157**	**27.83**	**0.9640**	**0.0342**	**27.68**	**0.9708**	**0.0217**	**29.05**	**0.9689**	**0.0263**
Anim-NeRF [6] (1 minute)	12.39	0.7929	0.3393	13.10	0.7705	0.3460	11.71	0.7797	0.3321	12.31	0.8089	0.3344
Ours (1 minute)	**29.65**	**0.9730**	**0.0192**	**27.97**	**0.9649**	**0.0346**	**27.90**	**0.9722**	**0.0249**	**28.92**	**0.9692**	**0.0180**

Table 1. **Qualitative Comparison with SoTA on the PeopleSnapshot [1] dataset.** We report PSNR, SSIM and LPIPS [68] between real images and the images generated by our method and two SoTA methods, Neural Body [49] and Anim-NeRF [6]. We compare all three methods at their convergence, and also compare ours with Anim-NeRF at 5 minutes, 3 minutes and 1 minute training time.

Occupancy-based regularization Previous methods for the learning of human avatars [6, 27] often encourage models to predict zero density for points outside of the surface and solid density for points inside the surface by leveraging the SMPL body model as regularizer. This is done to reduce artifacts near the body surface. However such regularization makes heavy assumptions about the shape of the body and does not generalize well for loose clothing. Moreover, we empirically found this regularization is not effective in removing artifacts near the body. This can be seen in Fig. 3. Instead of using SMPL for regularization, we use our occupancy grid which is a more conservative estimate of the shape of the subject and the clothing, and define an additional loss $\mathcal{L}_{\text{reg}}$ which encourages the points inside the empty cells of the occupancy grid to have zero density:

$$\mathcal{L}_{\text{reg}} = \begin{cases} |\sigma(\mathbf{x})| & \text{if } \mathbf{x} \text{ is in the empty space} \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

4. Experiments

We evaluate the accuracy and speed of our method on monocular videos and compare it with other SoTA methods. In addition, we provide an ablation study to investigate the effect of individual technical contributions.

Datasets

PeopleSnapshot We conduct experiments on the PeopleSnapshot [1] dataset, which contains videos of humans rotating in front of a camera. We follow the evaluation protocol defined in Anim-NeRF [6]. The pose parameters provided in this dataset are obtained using SMPLify [4], which do not always align with images. Hence, Anim-NeRF [6] optimizes the poses of training and test frames. For a fair quantitative comparison in Tab. 1, we train our model with the pose parameters optimized by Anim-NeRF and keep them frozen throughout training. Our model also supports body pose optimization. For all other results in the paper, we use an off-the-shelf 3D pose estimator and optimize poses jointly with our model to refine the pose estimates. This is done by back-propagating the gradient of the image reconstruction loss to the pose parameters. The camera parameters are given in PeopleSnapshot, obtained by standard calibration procedure.

SURREAL The PeopleSnapshot dataset has limited pose variations. To evaluate the performance on more challenging test poses, we also generate synthetic monocular sequences by rendering SMPL with texture maps from the SURREAL [58] dataset. For training, we drive the textured SMPL model with the same SMPL parameters from PeopleSnapshot, and for test, we generate challenging out-of-distribution poses. This allows us to evaluate the performance of methods on novel pose synthesis.

Baselines

Anim-NeRF [6] This baseline models human shapes and appearance in a canonical space with an MLP-based NeRF. Given a pose, they first generate an SMPL body in the target pose. Then for each query point in deformed space, its corresponding skinning weights are defined as the weighted average of skinning weights of its K nearest vertices on the posed SMPL mesh. Finally, with the skinning weights, the query point can be transformed back to the canonical space based on inverse LBS.

Neural Body [49] This baseline learns a set of latent codes anchored to a deformable SMPL mesh. These latent codes deform with the SMPL mesh and are decoded into radiance fields in different poses.

4.1. Comparison with SoTA

Reconstruction Quality To measure the appearance quality of the reconstructed avatar, we animate and ren-

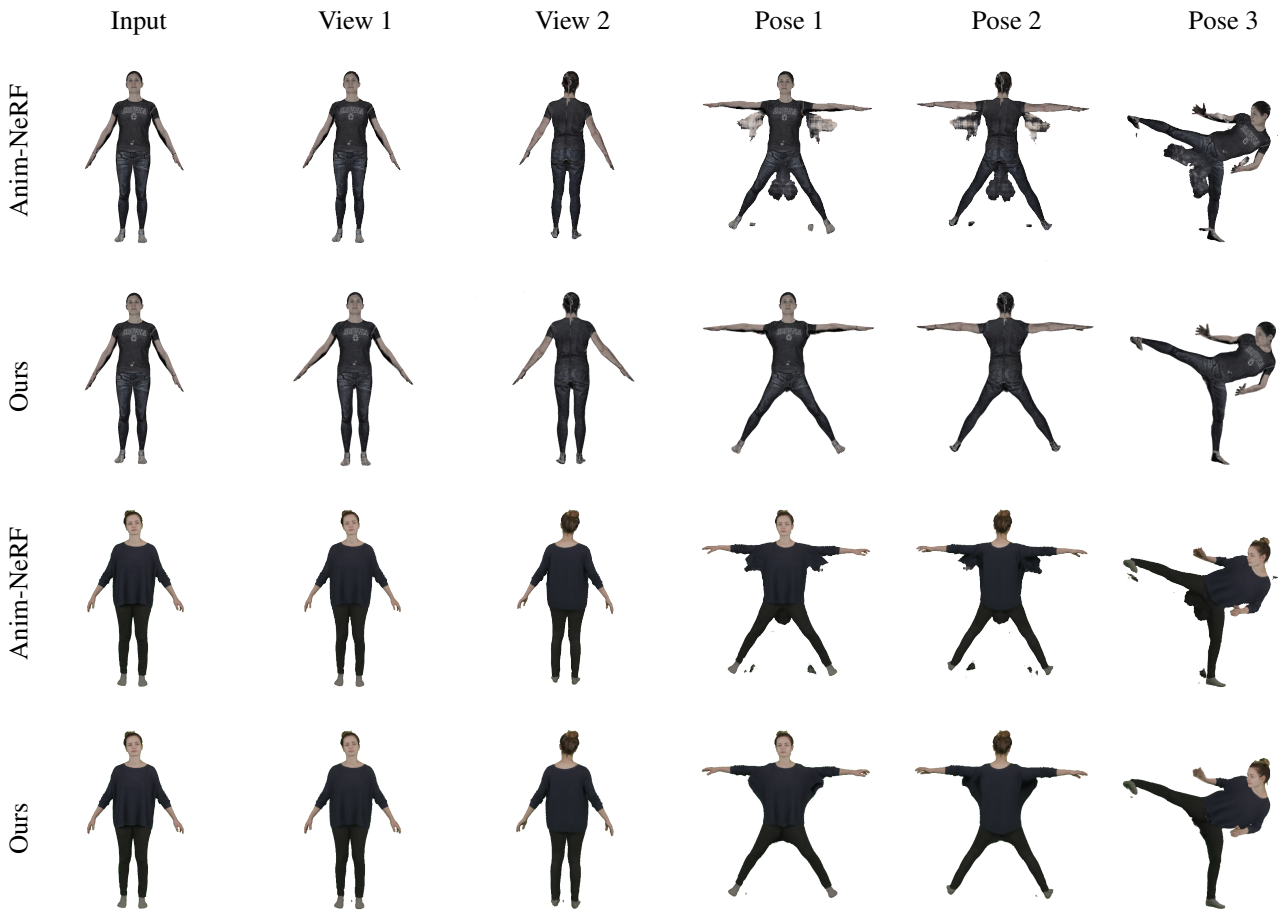

Figure 3. **Qualitative Results on SURREAL [58] and PeopleSnapshot dataset [1].** We show reconstructed avatars on SURREAL (top) and PeopleSnapshot (bottom) from different viewpoints (column 2-3) and in various poses (column 4-6).

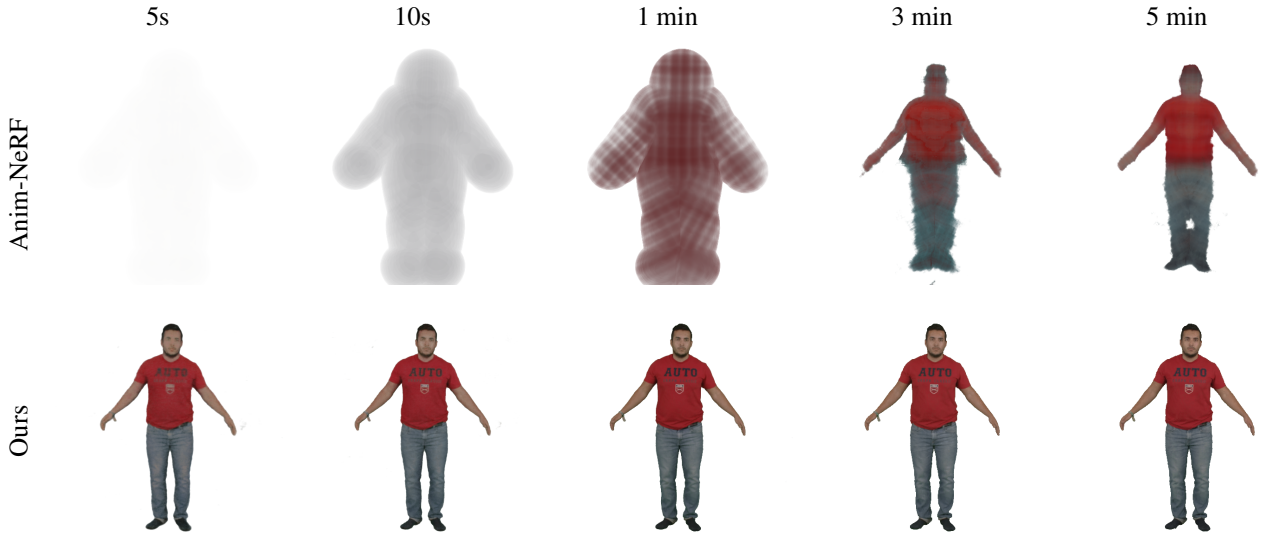

Figure 4. **Training Progression.** We show the image quality at different training iterations. Our method converges significantly faster than SoTA Anim-NeRF [6].

	Anim-NeRF			Ours		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
S1	21.66	**0.9450**	0.07615	**24.48**	0.9353	**0.0304**
S2	20.00	**0.9483**	0.09693	**23.94**	0.9354	**0.0343**
S3	20.06	0.9326	0.07948	**25.08**	**0.9494**	**0.0275**

Table 2. **Quantitative Results on the SURREAL Dataset.** We evaluate novel pose synthesis quality of our method and Anim-NeRF [6] on 3 synthetic subjects.

	PSNR↑	SSIM↑	LPIPS↓	Training Time↓
w/o Skipping	28.29	0.9680	0.030	3m 00s
w/ Skipping	28.66	0.9699	0.025	1m 42s
Update Frequency=8	28.73	0.9700	0.027	1m 44s
Update Frequency=16	28.66	0.9699	0.025	1m 42s
Update Frequency=32	28.56	0.9694	0.026	1m 41s
Decay Rate=0.5	28.74	0.9704	0.026	1m 57s
Decay Rate=0.8	28.66	0.9699	0.025	1m 42s
Decay Rate=0.9	28.62	0.9695	0.023	2m 03s
Resolution=32	28.31	0.9690	0.026	2m 17s
Resolution=64	28.66	0.9699	0.025	1m 42s
Resolution=96	28.81	0.9705	0.026	1m 58s

Table 3. **Ablation: Empty Space Skipping.** We perform an ablation study over the hyperparameters of empty space skipping. For all the experiments we report the average over 4 sequences in PeopleSnapshot after 50 epochs.

	PSNR↑	SSIM↑	LPIPS↓
w/o occupancy-based regularizer	28.22	0.9680	0.0301
w/ occupancy-based regularizer	**28.64**	**0.9700**	**0.0240**

Table 4. **Ablation: Occupancy-based Regularizer.** We evaluate image quality averaged over the 4 PeopleSnapshot sequences. For both cases we train our model for 50 epochs.

der the reconstructed model with the poses of test frames in PeopleSnapshot, and measure the difference between generated and real images. When training all methods to convergence, our generated images are significantly better than Neural Body [49] and achieve on-par quality as SoTA method Anim-NeRF [6], as indicated by the image quality metrics in Tab. 1 and the qualitative results in Fig. 3.

Speed Our method requires much less training time and computation resources than SoTA methods. We only require 1 minute to train on a single RTX 3090 while Anim-NeRF [6] requires 13 hours on $2\times$ RTX 3090 and Neural Body [49] requires 14 hours on $4\times$ RTX 2080. Ours also achieves superior rendering speed - we can render images at 540×540 resolution on a single RTX 3090 at more than 15 FPS, which is orders of magnitude faster than baselines.

Given the same training time budget, our method achieves significantly better image quality than Anim-NeRF as shown in Tab. 1. Comparing our training progression with Anim-NeRF in Fig. 4, we note that our method already learns meaningful appearance and moderate details within 5s and acceptable visual quality at 10s. After only 1 minute of training time, our method already achieves high-fidelity reconstruction quality. In contrast, Anim-NeRF does not produce meaningful results this early in training and only learns the rough shape after 3 minutes.

Figure 5. **Effect of Occupancy-based Regularization.** Without our regularization loss, the model suffers from floating artifacts. Our occupancy-based regularization loss successfully removes such artifacts. While a global sparsity prior biasing all densities towards 0 can also reduce such artifacts, it leads to degenerated image quality (semi-transparent).

Novel Pose Synthesis Quality The previous evaluation does not reflect the performance of novel pose synthesis, because the pose variation in the PeopleSnapshot dataset is limited (self-rotating). Due to the lack of ground truth images in novel poses, we resort to evaluating novel pose synthesis qualitatively. We generate images in novel challenging poses with our method and Anim-NeRF. As shown in Fig. 3, our method can faithfully generate images even in challenging body poses while preserving high fidelity. In contrast, Anim-NeRF suffers from artifacts under arms and between legs, because their methods cannot correctly disambiguate body parts that are close to each other in the posed space. Our method outperforms our baseline especially for loose clothing as shown in the bottom example in Fig. 3. This is because we don't rely on the SMPL body model for regularization and hence can better deal with subjects and clothing that differ from SMPL. To quantitatively evaluate novel pose synthesis, we generate synthetic data in challenging poses as ground truth. The results in Tab. 2 and Fig. 3 verify the superiority of our method in terms of novel pose synthesis quality.

4.2. Ablation Study

Empty Space Skipping We study the effect of our proposed empty space skipping scheme for dynamic objects. As shown in Tab. 3, skipping empty space significantly improves the training and rendering speed, and is robust to the choice of hyperparameters.

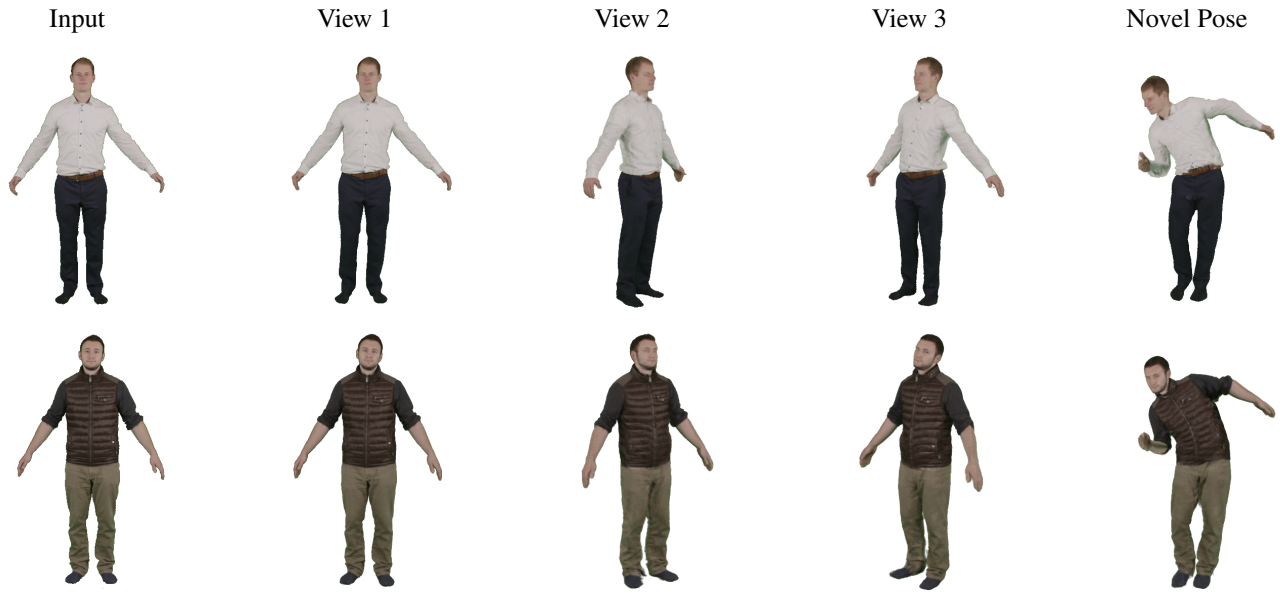

Figure 6. **More Qualitative Results of Our Method.** We show our reconstructed avatars from different views and in different poses.

Occupancy-based Regularization $\mathcal{L}_{\text{reg}}$ The occupancy grid for empty space skipping can also help regularize the radiance field to reduce noise via our regularization loss $\mathcal{L}_{\text{reg}}$ described in Section. 3.4. As shown in Fig. 5, this loss effectively reduces floating artifacts and consequently helps to improve the overall image quality as evidenced by the PSNR improvement in Tab. 4. Another common approach to reducing floating noise is to encourage zero density for every point in space. We compare our solution with this strategy and find it (Global Sparsity) leads to degenerated image quality as shown in Fig. 5.

More qualitative results are shown in Fig. 6

4.3. Limitations

Although occupancy-based regularization is generally effective at reducing floating artifacts, we occasionally observe remaining artifacts when the pose parameters are noisy. As illustrated in Fig. 7, these artifacts stem from the model compensating for noisy feet estimates to satisfy the image reconstruction loss (see below) and can't be removed via regularization.

Figure 7. **Limitation: Remaining Artifacts due to Inaccurate Pose.** The floating artifact between the feet (right) is generated to minimize the reconstruction loss in a different view (left).

Our method does not model facial expression and hand articulation, hence the reconstructed face and hand quality might degrade if facial expression or hand pose changes drastically. This could be addressed by introducing more complicated body models such as SMPL-X [47]. In addition, our method does not model pose-dependent deformations and view-dependent appearance changes, hence it's unable to model wrinkles and non-Lambertian objects such as eyeglasses. Finally, our method reconstructs avatars purely based on image observations and cannot infer unseen regions. For instance, if the input video only captures the front side of the subject, our method cannot reconstruct the back side. This limitation could potentially be addressed by leveraging learning-based methods to predict the texture and geometry of unobserved regions.

5. Conclusion

In this paper, we propose a method that can reconstruct animatable human avatars from monocular videos within 60 seconds and can animate and render the model afterward at 15 FPS. To achieve this, we combine an efficient neural representation, Instant-NGP [42], and an efficient articulation module Fast-SNARF [7]. This naive combination does not yield optimal speed. We devise an empty space skipping scheme to improve our rendering speed, and an occupancy-aware regularization loss to reduce floating artifacts in space. In comparison with SoTA methods, our method achieves on-par image quality while being significantly faster during training and inference. While this paper focuses on full-body human reconstruction, the idea could be applied to other objects. An interesting next step is to extend our method to reconstruct general articulated objects or animals from images efficiently.

Acknowledgements Xu Chen was supported by the Max Planck ETH Center for Learning Systems.

References

[1] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *Proc. of the International Conf. on 3D Vision (3DV)*, 2018. 1, 2, 5, 6

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8387–8397, Jun 2018. CVPR Spotlight Paper. 1, 2

[3] Alexander W. Bergman, Petr Kellnhofer, Wang Yifan, Eric R. Chan, David B. Lindell, and Gordon Wetzstein. Generative neural articulated radiance fields. *Arxiv*, 2022. 2

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2016. 5

[5] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2022. 2

[6] Jianchuan Chen, Ying Zhang, Di Kang, Xuefei Zhe, Linchao Bao, Xu Jia, and Huchuan Lu. Animatable neural radiance fields from monocular rgb videos. *arXiv.org*, 2021. 1, 2, 5, 6, 7

[7] Xu Chen, Tianjian Jiang, Jie Song, Max Rietmann, Andreas Geiger, Michael J. Black, and Otmar Hilliges. Fast-SNARF: A fast deformer for articulated neural fields. *arXiv*, abs/2211.15601, 2022. 2, 3, 8

[8] Xu Chen, Tianjian Jiang, Jie Song, Jinlong Yang, Michael J Black, Andreas Geiger, and Otmar Hilliges. gdna: Towards generative detailed neural avatars. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[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 *International Conference on Computer Vision (ICCV)*, 2021. 2, 3, 4

[10] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. SNARF: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 2

[11] Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. Implicit functions in feature space for 3D shape reconstruction and completion. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[12] 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 Trans. on Graphics*, 34, 2015. 2

[13] Boyang Deng, JP Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Neural articulated shape approximation. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[14] Zijian Dong, Chen Guo, Jie Song, Xu Chen, Andreas Geiger, and Otmar Hilliges. PINA: Learning a personalized implicit neural avatar from a single RGB-D video sequence. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[15] 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 Transactions on Graphics (ToG)*, 35(4):1–13, 2016. 2

[16] 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

[17] Chen Guo, Xu Chen, Jie Song, and Otmar Hilliges. Human performance capture from monocular video in the wild. In *2021 International Conference on 3D Vision (3DV)*, pages 889–898. IEEE, 2021. 2

[18] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. *arXiv*, 2023. 2

[19] 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 Trans. on Graphics*, 38, 2019. 2

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

[21] Marc Habermann, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2020. 2

[22] Tong He, John Collomosse, Hailin Jin, and Stefano Soatto. Geo-PIFu: Geometry and pixel aligned implicit functions for single-view human reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[23] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. ARCH++: Animation-ready clothed human reconstruction revisited. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[24] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *ACM Trans. Gr.*, 2022. 2

[25] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[26] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[27] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field

from a single video. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2022. 2, 5

[28] Yue Jiang, Marc Habermann, Vladislav Golyanik, and Christian Theobalt. Hifecap: Monocular high-fidelity and expressive capture of human performances. In *BMVC*, 2022. 2

[29] Ruilong Li, Matthew Tancik, and Angjoo Kanazawa. Nerfacc: A general nerf accleration toolbox. *arXiv preprint arXiv:2210.04847*, 2022. 2, 3

[30] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhoefer, Jürgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2022. 2

[31] Siyou Lin, Hongwen Zhang, Zerong Zheng, Ruizhi Shao, and Yebin Liu. Learning implicit templates for point-based clothed human modeling. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2022. 2

[32] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2, 3

[33] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *NeurIPS*, 2020. 2, 3

[34] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural Actor: Neural free-view synthesis of human actors with pose control. *ACM Trans. on Graphics*, 2021. 2

[35] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. on Graphics*, 2015. 2, 3

[36] Wojciech Matusik, Chris Buehler, Ramesh Raskar, Steven J Gortler, and Leonard McMillan. Image-based visual hulls. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, 2000. 2

[37] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3D reconstruction in function space. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

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

[39] Marko Mihajlovic, Yan Zhang, Michael J Black, and Siyu Tang. LEAP: Learning articulated occupancy of people. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[40] 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, 4

[41] 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 *Proc. of the European Conf. on Computer Vision (ECCV)*, 2020. 2

[42] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, July 2022. 2, 3, 8

[43] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 2

[44] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Unsupervised learning of efficient geometry-aware neural articulated representations. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2022. 2

[45] Michael Oechsle, Lars Mescheder, Michael Niemeyer, Thilo Strauss, and Andreas Geiger. Texture fields: Learning texture representations in function space. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019. 2

[46] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[47] 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 *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 8

[48] 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 *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 2

[49] 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 *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 5, 7

[50] Daniel Rebain, Mark Matthews, Kwang Moo Yi, Dmitry Lagun, and Andrea Tagliasacchi. Lolnerf: Learn from one look, 2022. 4

[51] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 2

[52] 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 *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019. 2

[53] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[54] Sara Fridovich-Keil and Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, 2022. 2, 3

[55] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[56] Towaki Takikawa, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, and Sanja Fidler. Neural geometric level of detail: Real-time rendering with implicit 3D shapes. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[57] Garvita Tiwari, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Neural-GIF: Neural generalized implicit functions for animating people in clothing. In *International Conference on Computer Vision (ICCV)*, October 2021. 2

[58] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 5, 6

[59] Shaofei Wang, Andreas Geiger, and Siyu Tang. Locally aware piecewise transformation fields for 3D human mesh registration. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[60] Shaofei Wang, Marko Mihajlovic, Qianli Ma, Andreas Geiger, and Siyu Tang. MetaAvatar: Learning animatable clothed human models from few depth images. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[61] Shaofei Wang, Katja Schwarz, Andreas Geiger, and Siyu Tang. Arah: Animatable volume rendering of articulated human sdfs. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2022. 2

[62] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220, June 2022. 1, 2

[63] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. ICON: Implicit Clothed humans Obtained from Normals. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[64] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-NeRF: Neural radiance fields for rendering and temporal reconstruction of humans in motion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[65] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. Monoperfcap: Human performance capture from monocular video. *ACM Trans. Graph.*, 37(2), 2018. 2

[66] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 2, 3

[67] Jianfeng Zhang, Zihang Jiang, Dingdong Yang, Hongyi Xu, Yichun Shi, Guoxian Song, Zhongcong Xu, Xinchao Wang, and Jiashi Feng. Avatargen: A 3d generative model for animatable human avatars. *Arxiv*, 2022. 2

[68] 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. 5

[69] Hao Zhao, Jinsong Zhang, Yu-Kun Lai, Zerong Zheng, Yingdi Xie, Yebin Liu, and Kun Li. High-fidelity human avatars from a single rgb camera. In *CVPR*, 2022. 2

[70] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. PaMIR: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 2021. 2