HumanNeRF-SE: A Simple yet Effective Approach to Animate HumanNeRF with Diverse Poses

Caoyuan Ma[1] Yu-Lun Liu[2] Zhixiang Wang[3,4] Wu Liu[5] Xinchen Liu[6] Zheng Wang[1†]
[1]National Engineering Research Center for Multimedia Software, Institute of Artificial Intelligence, School of Computer Science, Wuhan University [2]National Yang Ming Chiao Tung University [3]The University of Tokyo
[4]National Institute of Informatics [5]School of Information Science and Technology, University of Science and Technology of China [6]JD Explore Academy

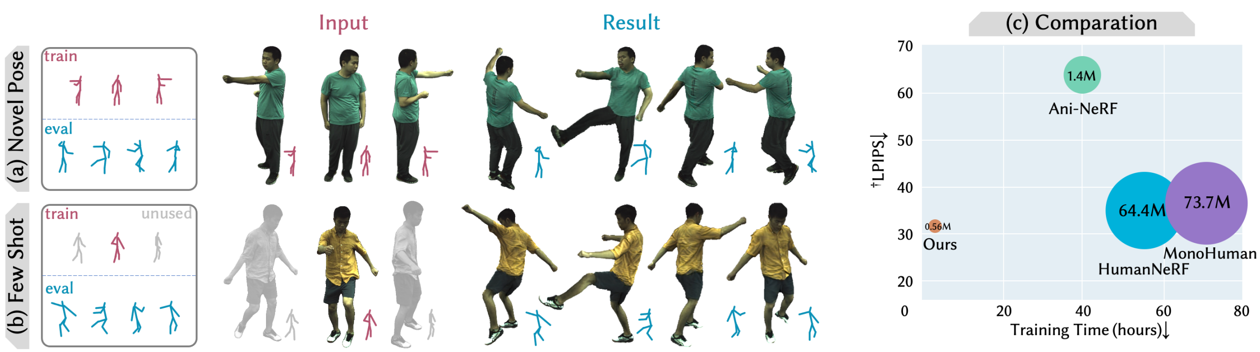


Figure 1. **Overview.** HumanNeRF-SE efficiently synthesizes images of performers in *diverse* poses, blending *simplicity* with *effectiveness*. It outperforms previous methods by creating a wider range of new poses (a), maintains generalization without overfitting with limited input frames (b), and requires fewer than 1% of learnable parameters, reducing training time by 95% while delivering superior results in the few-shot scenario (c). †LPIPS = 1,000×LPIPS. Project page: https://miles629.github.io/humanNeRF-se.github.io/

Abstract

We present HumanNeRF-SE, a simple yet effective method that synthesizes diverse novel pose images with simple input. Previous HumanNeRF works require a large number of optimizable parameters to fit the human images. Instead, we reload these approaches by combining explicit and implicit human representations to design both generalized rigid deformation and specific non-rigid deformation. Our key insight is that explicit shape can reduce the sampling points used to fit implicit representation, and frozen blending weights from SMPL constructing a generalized rigid deformation can effectively avoid overfitting and improve pose generalization performance. Our architecture involving both explicit and implicit representation is simple yet effective. Experiments demonstrate our model can synthesize images under arbitrary poses with few-shot input and increase the speed of synthesizing images by 15 times through a reduction in computational complexity without using any existing acceleration modules. Compared to the state-of-the-art HumanNeRF studies, HumanNeRF-SE achieves better performance with fewer learnable parameters and less training time.

1. Introduction

Neural Radiance Field (NeRF) [33, 54, 55, 58, 63] represents the scene as an implicit field and utilizes volumetric renderer to synthesize the scene has demonstrated remarkable advancements in reconstruction and *novel-view* syntheses of *static* scenes. However, they typically do not account for object deformation and perform poorly on *dynamic* humans due to the complex deformation caused by motions. Deformable NeRFs endow implicit fields with the capability to express dynamic objects [37, 38, 42, 53, 59] or even humans [11, 16, 17, 56, 62]. Although these methods [11, 16, 17, 56, 62] could learn high-quality human representations and synthesize images from arbitrary viewpoints, they cannot synthesize images with *novel poses* that are significantly different from that of the training videos.

We aim to automatically render photo-realistic human images with arbitrary *viewpoints* and *poses* from monocular videos. Although there are some studies [5, 18, 26, 40, 41, 66] have attempted to learn animatable human representations by introducing neural blend weights or using UV-referenced coordinate systems, their requirement for multi-camera data limits their practical applicability. The problem becomes practical with monocular inputs but highly *ill-*

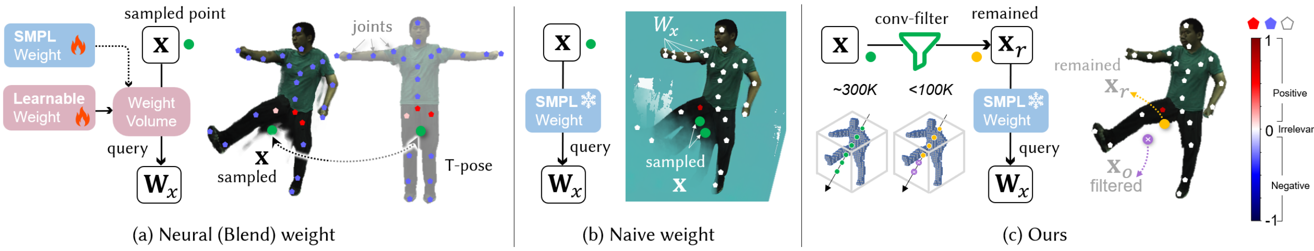


Figure 2. **Different weights for deformation.** (a) Prior methods [26, 40, 56, 62] learn a weight volume for deformation through neural networks or fine-tune blending weights obtained from fitting SMPL to the input frame. The weight volume optimized along with NeRF parameters per human image is prone to over-fitting. When synthesizing novel pose images, the over-fitted weights will deform points onto the canonical space *incorrectly* and lead to artifacts. (b) Our idea is to use SMPL's blending weights directly because these weights are pre-trained on numerous human images to avoid overfitting. However, simply utilizing the nearest SMPL vertex's blending weights for deformation fills the sampling space with incorrect colors as the training phase deforms irrelevant sampling points onto the human body. (c) We propose to filter irrelevant points according to the human body information of SMPL. This way, we can avoid over-fitting and reduce the number of sampling points.

posed due to the limited and patchy observations.

Existing monocular-based methods [11, 16, 17, 56, 62] usually decompose the human implicit field into rigid and non-rigid components, reducing the ill-posedness in joint optimization. These two components deform sampling points from the observation space to the canonical one. The non-rigid field is learned by a neural network conditioned on the human pose or frame index. On the contrary, the rigid field uses an explicit model—Linear Blending Skinning (LBS)—given blending weights learned from scratch or fine-tuning SMPL's [29] weights. Since the data for training NeRF is limited and the number of optimizable parameters is large, the blending weights could overfit the input data and yield unsatisfactory results, especially when the input poses become very restricted (Figure 2a).

In this paper, we present HumanNeRF-SE, which synthesizes novel pose images with *tens* of simple inputs and *a few* learnable parameters in *hours*. Our approach distinguishes itself from previous methods like HumanNeRF [56] by effectively leveraging prior knowledge provided by SMPL. On the one hand, we use the blending weights from pre-trained SMPL *without* any change for rigid deformation. This is because SMPL trained on numerous human data is generalizable to diverse humans and poses. On the other hand, we employ the SMPL's vertices for sampling points. The motivation is that we found simply using the blending weights is not enough (Figure 2b) since there are a lot of irrelevant human points in the volume. These points could be deformed to the human body to be reconstructed incorrectly. We propose the Conv-Filter guided by SMPL's vertices to reduce the irrelevant points (Figure 2c). Our method not only avoids overfitting but also greatly reduces the required sampling points from $\sim$300K to less than 100K. Besides sampling, we also use spatial-aware features extracted from SMPL's vertices to condition non-rigid deformation.

Specifically, we first voxelize SMPL's vertices by employing a sparse convolution to diffuse the vertices across the voxel volume. Second, the spatial-aware feature and occupancy of the sampling point can be easily queried in this volume. Throughout this procedure, a significant number of points unrelated to the body are filtered out, leaving behind only those points likely to be related to the human body. Third, these points can be deformed to a unified canonical pose using the rigid deformation to get the coarse coordinates. Fourth, we refine the coarse coordinates by performing a non-rigid mapping conditioned on the point-level spatial-aware feature obtained in the convoluted voxel volume. Finally, we get the colors and densities of the sampling points through a neural network and render the image through a differentiable renderer.

Overall, we propose a *simple* yet *effective* HumanNeRF method that synthesizes images of varying poses *efficiently*. Our approach utilizes explicit SMPL prior knowledge to design a generalized rigid deformation and a specific non-rigid deformation to map points from observation space to canonical space. Our experiments show that our method can generate novel poses with significant differences from the training poses, even when the input is limited to a few shots. We also further demonstrate the superiority of our method on our captured in-the-wild data, where the input video involves a simple rotation of the user.

To sum up, we make the following contributions:

- We design modules to leverage the prior knowledge from SMPL for deformations and point sampling. This dramatically reduces computational complexity and avoids overfitting.
- Our architecture is simple yet effective. Compared to methods with similar performance, HumanNeRF-SE uses less than 1% learnable parameters, 1/20 training time, and increases rendering speed by 15 times without using any existing acceleration modules.

2. Related Work

3D Performance Capture. Recently, deep neural networks are commonly used to learn scene or human representation from images, with a range of methods now available including voxels [28, 45], point clouds [1, 12, 57, 61], textured meshes [15, 23, 24, 43, 44, 52, 60, 68, 69], multiplane images [8, 70], and implicit functions [25, 27, 33, 36, 42, 46]. Most of these methods aim to optimize a detailed 3D geometry, and then synthesize results into images or videos for application, which limited their usage. Our method can directly synthesize human body images.

Neural Rendering. To synthesize novel view images of a static object without recovering detailed 3D geometry, previous neural rendering method [2, 13, 19, 30, 31, 35, 47, 51, 54, 55, 58, 63, 65] represent amazing image synthesis result, but these methods typically assume multi-camera input and usually don't take object's deformation into consider. The deformable methods [9, 21, 37, 38, 42, 53, 59] endow the implicit field with the ability to express dynamic objects but don't perform well in the human because of the complexity of human deformation. Our method employs a simple and effective architecture for human body deformation. The ability to synthesize the movements of specific human bodies in different poses has a wide range of applications. Therefore, it is very meaningful to extend methods to adapt to dynamic human bodies.

Multi-camera HumanNeRF. There has been some research [4–6, 10, 14, 18, 22, 26, 40, 41, 66] on learning dynamic human representation through NeRF from multi-camera images. NeuralBody [41] uses structured pose features generated from SMPL [29] vertices to anchor sampling points in any poses from sparse multi-camera videos, which inspired us to use spatial information of vertices. [10, 14, 22, 26, 40, 66] transform sampling points of dynamic human to a canonical space for NeRF training. Because the information in monocular data is much more limited than in multi-camera data, some of them have the ability to train with only monocular video, but they are not designed for monocular scenes and usually don't perform well in monocular data. Our method only requires few-shot input and also performs well in monocular data.

Monocular HumanNeRF. Since obtaining monocular videos is much easier than obtaining synchronized multi-view videos, it is significant to extend the capabilities of HumanNeRF to monocular videos. Inspired by [37, 38, 42], which maps rays in dynamic scenes to canonical space, [11, 16, 17, 56, 62] introduce priors to regularize the deformation. [11, 16, 39] greatly improve the speed of training and rendering by using more efficient spatial encoding methods [3, 34], while these encoding methods don't perform well in terms of acceleration in our experiments. [48] learn dynamic human bodies by modeling the relationship between sampling points and joints. [17, 40] learns the blending field by extending the deformation weight of the closest correspondence from the SMPL mesh, which leads to a significant increase in computational cost. HumanNeRF [56] demonstrates amazing novel view results by decoupling the motion field, which uses a significant amount of neural networks to fit various modules and often takes a very long time to train. Monohuman [62] further improves the pose generalization performance of HumanNeRF by adding a reference image module and consist loss. Our experiments aimed to explore the fundamental reasons why HumanNeRF performs poorly in pose generalization, and to achieve better results on more challenging data by rebuilding a simple yet efficient architecture with SMPL vertices to combine explicit and implicit human representation.

3. HumanNeRF-SE

We propose a simple yet effective approach for learning the implicit representation of human bodies from limited inputs and being capable of synthesizing diverse novel poses. Compared to other similar methods, each module of our approach is designed to map the explicit and implicit human representation better and generalize to arbitrary novel poses without overfitting. This results in fewer data demands and computational load and improves pose generalization capabilities compared to other approaches. The pipeline is illustrated in Figure 3.

The key to learning the deformable NeRF representation of human bodies lies in canonicalizing the sampling points within the dynamic observation space. Prior methods predominantly depend on neural network fitting or supplementary texture information to precisely anchor the sampling points relative to the human body. These methods also introduce frame-level features to augment the multi-view results, albeit at the expense of substantially diminishing the pose generalization capabilities of the model.

Different from these methods, our method effectively leverages the explicit vertices $\mathbf{v}$ to canonicalize the sampling points $\mathbf{x}$ within the dynamic human observation space. This process involves constructing voxel volume $\mathbf{V}$ (Sec. 3.1), convoluting voxel volume channel-by-channel to obtain spatially-aware features F_s and filtering out useless sampling points $\mathbf{x}_o$ and get human-related points $\mathbf{x}_r$ (Sec. 3.2), mapping points to canonical space generally D and specifically P (Sec. 3.3). It is important to emphasize that we exclusively utilize point-level features throughout the entire process, abstaining from the use of any frame-level features, thereby ensuring the pose generalization capabilities of our model.

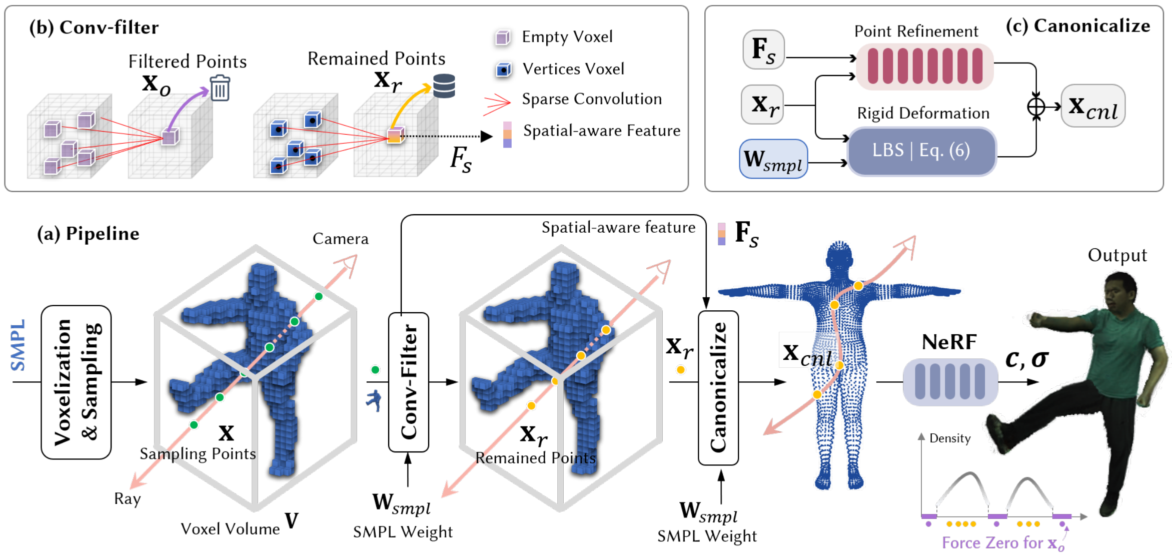


Figure 3. **Framework of HumanNeRF-SE.** (a) We first voxelize the observation space as a voxel volume $\mathbf{V}$. For a voxel containing vertices, the value will be the number of vertices (as one occupancy channel) and the corresponding SMPL weight. (b) We performed channel-by-channel convolution on the volume. All sampling points are queried in the convolutional volume to get their spatial-aware features. Those points with zero occupancy will be filtered out. (c) We query the nearest weight of the remained points in the volume, which is used for rigid deformation. Spatial-aware features are utilized in the neural network to correct the rigid results and obtain the final point coordinates in the canonical space. The sampling points in the canonical space obtain their colors and densities through the NeRF network. The densities of filtered points are forced to be zero.

In summary, our method can be represented as:

$$\mathbf{c}, \sigma = M_\sigma(P(\mathbf{F}_s, \mathbf{x}_r) + D(\mathbf{x}_r, K(\mathbf{x}_r, \mathbf{V}), J)) \quad (1)$$

where J represents the pose and M_σ is the NeRF network which is similar to baselines for better comparison. We use K to query the nearest weights of points $\mathbf{x}$ in the volume $\mathbf{V}$. Our rigid deformation D is a general mapping process, and it is not influenced by the training individual.

3.1. Voxelization

In order to more efficiently handle the relationship between the sampling points and the SMPL model, we first voxelized the SMPL space. However, since the vertices of the SMPL model within the boundary of the human body are still relatively sparse, and we used the Sparse Convolution Sp [7] to construct our voxel volume $\mathbf{V}$.

We processed data similarly to NeuralBody [41] in this part. For a given set of SMPL vertices $\mathbf{v}$, we first calculate the maximum and minimum values on the coordinate axes to get the bounding box, scale the bounding box to the set voxel size $\mathbf{vs}$, and find the least common multiple axes and 32, in preparation for subsequent sparse convolution. For each SMPL vertex, we also scale it down according to voxel size after subtracting the minimum value.

$$V = Sp(\mathbf{v}, \mathbf{vs}, \mathbf{W}) \quad (2)$$

In the generated voxel volume, each voxel that contains vertices holds two values: the corresponding LBS weight w_j in SMPL weight $\mathbf{W}_{smpl}$ and the count of contained vertices n_v as an occupancy indicator. For voxels without any contained vertices, all channels are assigned a value of zero. The value of a certain voxel is:

$$\begin{cases} (n_v, w_1, w_2 \ldots, w_{24}), & \text{if contain vertices} \\ (0, 0, 0 \ldots, 0), & \text{if empty voxel} \end{cases} \quad (3)$$

3.2. Conv-Filter

We innovatively use spatial convolution to filter the sampling points and extract features simultaneously. A convolution kernel is initialized to one for convolving the value of the voxel volume. To preserve high-frequency information, we use channel-by-channel convolution.

$$F_{s_i} = \sum_{m=h'}^{h} \sum_{n=w'}^{w} \sum_{t=d'}^{d} \nu_{m,n,t,i} \cdot \vartheta_{m,n,t,i} \quad (4)$$

where $h' = h - k$, $w' = w - k$ and $d' = d - k$. The i-th channel of F_s results from convolving i-th channel of voxel's value ν and kernel's weight ϑ. If the occupancy of the convolution result is zero, it means that the current convolution center coordinate is not related to the human body (Eq. 3). We force these points not to participate in subsequent calculations and set their density value d to zero.

$$\begin{cases} \mathbf{x}_o \text{ filter out}, d(\mathbf{x}_o) = 0 & \text{if } \mathbf{F}_s(x) = 0 \\ \mathbf{x}_r \text{ remain} & \text{if } \mathbf{F}_s(x) > 0 \end{cases} \quad (5)$$

3.3. Point Canonicalization

We compute the rotation R_j and translation T_j of the current pose relative to the joints in the T-pose of the human body in a similar way of HumanNeRF [56]. Inspired by NeuralBody [41]'s use of convolution to diffuse vertex occupancy, we use a simpler method to query the nearest neighbor SMPL weight $\mathbf{W}_{smpl} = K(\mathbf{x}, \mathbf{v})$

$$\mathbf{x}^r_{cnl} = \sum_{j \in J} w_j \left((\mathbf{x}_r \cdot R_j) + T_j \right) \tag{6}$$

where J denotes a set of joints, w_j is the deformation weight of the j-th channel of $\mathbf{W}_{smpl}$ on the current sampled point, and $\mathbf{x}^r_{cnl}$ represents the coordinates of the sampled point in the canonical space after rigid deformation.

Directly using the nearest weight amounts to giving up non-rigid modeling of the possible clothing deformation caused by human body deformation, which often leads to unsatisfactory results. To enable the network to learn non-rigid deformation from limited images as much as possible, some methods [41, 56, 62] introduce *frame-level* features to facilitate learning. These features may include frame index or body pose, and for a particular frame, the frame-level features of all sampling points are generally consistent. This method is useful in the task of synthesizing new viewpoints in dynamic scenes, but it is not suitable for *animatable* human body. We designed a new network here to refine the coordinates of the sampled points in the canonical space. We use limited *point-level* features to learn the offset of the sampling points in canonical space to improve the novel-view evaluation metric. The spatial-aware feature has shown significant advantages in this process because it aggregates vertex information within the receptive field of the sampling point in the current pose:

$$\mathbf{x}_{cnl} = \mathbf{x}^r_{cnl} + P(\mathbf{F}_s, \mathbf{x}_r) \tag{7}$$

where $\mathbf{x}_{cnl}$ is the optimized result of the sampled point coordinates in the canonical space, and P is the network module for optimizing the sampled point coordinates in the canonical space.

3.4. Appearance Net and Rendering

In order to better compare with methods such as HumanNeRF [56], which rely on neural networks to fit the deformation weights, we used the same neural radiance field structure. The coordinates of the sampled points in the canonical space were encoded using the same positional encoding as in NeRF, and the MLP was used to output the corresponding colors and densities:

$$\mathbf{c}, \sigma = M_\sigma(\mathbf{x}_{cnl}) \tag{8}$$

Finally, we render the neural human by the volume renderer [32]. The rendered color $C(\mathbf{r})$ of the corresponding pixel with N_s samples of ray $\mathbf{r}$ can be written as:

$$\tilde{C}_t(\mathbf{r}) = \sum_{k=1}^{N_s} \left(\prod_{j-1}^{i-1} (1 - \alpha_j) \right) \alpha_i \, \mathbf{c}(x_i) \tag{9}$$

$$\text{where } \alpha_i = 1 - \exp(-\sigma(\mathbf{x}_i)\Delta t_i) \tag{10}$$

3.5. Training

Given a set of monocular videos, the frame images of the video $\{I_n | n = 1, 2, ..., N\}$. Most of the images are used for training and the rest are used for evaluation. The foreground mask Θ_{fore} is obtained from the density after the network output. For each foreground ray $\mathbf{r} \in \mathcal{R}$, the corresponding loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{LPIPS}} + \lambda \mathcal{L}_{\text{MSE}} \tag{11}$$

$$\text{where } \mathcal{L}_{\text{MSE}} = \frac{1}{\|\mathcal{R}\|} \sum_{\mathbf{r} \in \mathcal{R}} \left\| \tilde{C}(\mathbf{r}) - C(\mathbf{r}) \right\|_2^2 \tag{12}$$

$$\text{and } \mathcal{L}_{\text{LPIPS}} = lpips\left(\Theta_{\text{fore}}(\tilde{I}_i), \Theta_{\text{fore}}(I_i) \right) \tag{13}$$

4. Experiments

Evaluation Metrics. We use three metrics: peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and learned perceptual image patch similarity (LPIPS). It should be noted that LPIPS is the most *human-perceptually-aligned* metric among these indicators, while PSNR prefers smooth results but may have bad visual quality [64].

Dataset. We use ZJU-MoCap and our captured in-the-wild videos to evaluate our method. We follow [56] [62] to select the same six subjects in ZJU-MoCap for our evaluation. ZJU-MoCap is a dataset that captures the target human body from 23 different perspectives synchronously in a professional light stage room. We only use the first view captured in each subject.

The previous work did not consider the issue of pose leakage caused by the strong repetition of actions in ZJU-MoCap. In order to further validate the performance of the model and make the task more applicable, we shot videos using handheld devices. We limited the training videos to a person spinning one round and used diverse action videos for evaluation, which is a more real-world applicable evaluation method.

Competed Methods. We compared our method in terms of the performance of image synthesis with the most influential method HumanNeRF [56] and the latest state-of-the-art method MonoHuman [62] which improves the pose generalization of HumanNeRF and works better than [17, 41]. We also take Ani-NeRF [40] as one of the baselines because this work presents SMPL-based neural blend weight that can better generalize novel poses.

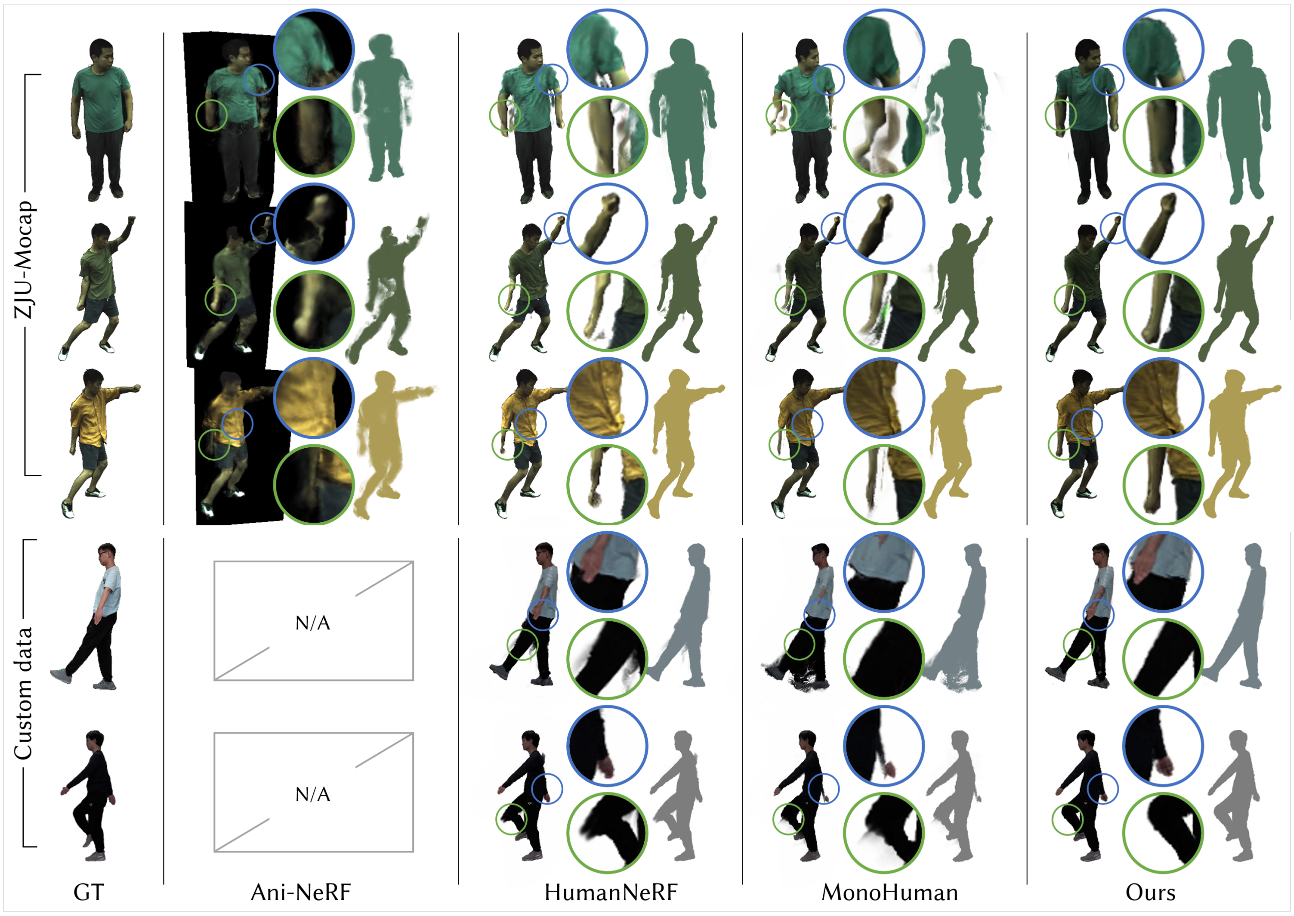


Figure 4. **Qualitative results with few-shot training images.** Because of limited information used in training, previous methods [40, 56, 62] cannot learn appropriate human weights. The official code of Ani-NeRF [40] did not produce reasonable results on our data since it is designed for multi-camera input. HumanNeRF [56] exhibits distortion and artifacts. The performance of Monohuman [62] is heavily influenced by the specific data.

Table 1. **Average results of six subjects on ZJU-MoCap**. Our method (blue) exhibits excellent quantitative metrics, especially in terms of LPIPS. This indicates that the results of our method are more in line with human visual perception. Our method demonstrated a better ability to avoid overfitting compared to other methods on our custom data. The official code of Ani-NeRF did not produce reasonable results on our custom data. †LPIPS = 1,000×LPIPS.

Methods	ZJU-MoCap						In-the-wild Data					
	Full			Few-shot			Full			Few-shot		
	PSNR↑	SSIM↑	†LPIPS↓	PSNR↑	SSIM↑	†LPIPS↓	PSNR↑	SSIM↑	†LPIPS↓	PSNR↑	SSIM↑	†LPIPS↓
Ani-NeRF [40]	21.24	0.8458	68.221	22.18	0.8339	64.839	–	–	–	–	–	–
HumanNeRF [56]	**31.15**	0.9739	24.822	29.90	0.9683	33.056	28.97	0.9629	48.128	28.82	0.9618	50.240
MonoHuman [62]	30.91	0.9718	31.292	30.10	0.9677	36.494	29.15	0.9639	51.623	29.21	0.9636	56.220
Ours	31.09	**0.9740**	**24.085**	**30.11**	**0.9684**	**32.084**	**29.23**	**0.9666**	**46.308**	**29.26**	**0.9669**	**47.161**

4.1. Quantitative Evaluation

For any set of data in ZJU-MoCap, we divided the data into training and testing data in a 4:1 ratio, which follows the setting of previous work. Differently, we uniformly select only about 30 frames from the divided training data as few-shot input. This setting can avoid pose leakage compared to the full input in a certain space. Our experimental results in ZJU-MoCap with few-shot training images are shown in Figure 4, and all the experimental results are shown in Table 1.

For our custom in-the-wild data, we train with a video of a performer spinning one round and use another video of the same person doing different poses for evaluation. The results are shown in Figure 4.

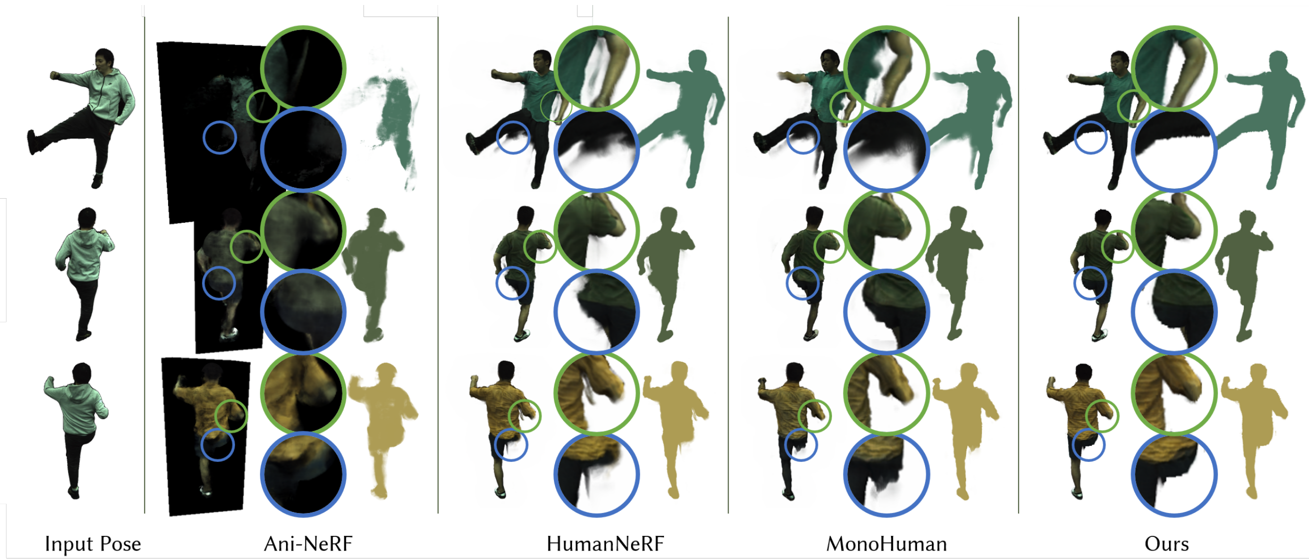
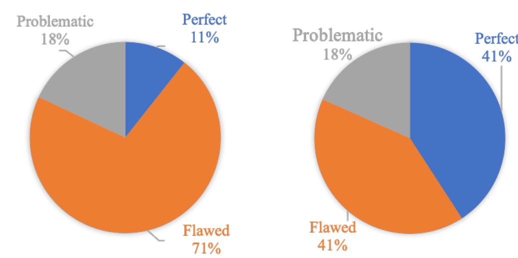


Figure 5. **Rendering results with pose sequences from Subject-387 in ZJU-MoCap.** We use all the videos of the performers to train and synthesize images with different pose sequences from Subject-387. The baselines produce noticeable artifacts, while our method maintains high-quality image synthesis.

Table 2. **Comparison of training and rendering time.** Our method does not use existing acceleration modules. The simple yet effective architecture greatly reduces the computation required, thereby improving the overall speed.

	Training	Rendering
Ani-NeRF [40]	40 h	2.73 s/it
HumanNeRF [56]	56 h	2.37 s/it
MonoHuman [62]	70 h	5.96 s/it
Ours	**2.5 h**	**0.16 s/it**



(a) HumanNeRF [56] (b) Ours

Figure 6. **Subjective evaluation results.** We use cross-rendering of different subjects in the ZJU-MoCap to ensure that the evaluation of the novel pose is sufficiently novel. Our results show a significantly higher perfect proportion than HumaneNeRF.

4.2. Qualitative Evaluation

In previous work, researchers often divide data into a certain ratio as novel pose experiments for quantitative evaluation. However, in the most widely used dataset ZJU-MoCap, performers repeat the same action in a set of data. This leads to the poses in evaluation data being highly similar to poses used for training. Although we mitigate pose leakage issues through few-shot input, a considerable portion of poses in the evaluation data are still similar, resulting in limited difference in the average results.

A simple way to avoid pose leakage problem is to use completely different action sequences as evaluation. But as there is no ground truth of ZJU-MoCap, we provided comparison results as Figure 5, and synthesized over 200 zero-shot pose images and converted them into 20 videos, which were subjectively evaluated by six or more participants as shown in Figure 6.

The entire evaluation process is single-blind, meaning that the participants do not know which specific method generated the results. We also included some test seeds, which serve as the Ground Truth, and all of these seeds received high scores from the participants, indicating that their evaluations were professional and objective. The participants were asked to evaluate the video in terms of image distortion, artifacts, details, plausibility, and precision, and to provide a final score that was assigned to one of three different levels.

As a result, almost all the participants subjectively think our results have a better performance. Our results were classified as perfect in significantly higher numbers than HumanNeRF, with fewer votes for flaws. However, both the results performed poorly in cases of poor data quality.

4.3. Ablation Studies

Conv-Filter. In our method, the point filter is essential. Our experiments showed that if the sampled points are not

w/ Conv-Filter *w/o* Conv-Filter

Figure 7. **Ablation study on Conv-Filter.**

w/ Point Refinement *w/o* Point Refinement

Figure 8. **Ablation study on Point Refinement.**

Point-level feature (Ours) Frame-level feature

Figure 9. **Ablation study on the input of Point Refinement**.

filtered, we cannot learn the correct alpha map (see Figure 7) in our experiments, and it greatly increases the computational complexity, requiring longer training time. We further investigated the reason why the phenomenon of alpha map learning errors occurs due to color diffusion into the surrounding space, and we believe that this is determined by the distribution of skin weight. As shown in Figure 2, the learnable weights tend to give negative values to irrelevant joint weights, but this is unreasonable. It can be explained by the fact that these methods do not require filters to avoid the phenomenon, because we observe similar phenomena when we map the learnable weights to the same distribution as ours through sigmoid.

Point-level Feature Refine. It is a common practice to add offsets to the deformation process using neural networks, but previous methods often used time or the pose of the current frame as control information. This frame-level feature often leads to overfitting, but in previous experiments, this phenomenon was not significant due to the similarity and repetition of actions in ZJU-MoCap. We extract point-level spatial-aware feature while filtering points in ConvFilter. This not only corrects the unnatural joints caused by rigid deformation (see Figure 8, Table 3) but also avoids overfitting compared to the previous frame-level feature (see Figure 9).

Table 3. **Quantitative ablation study in ZJU-MoCap.** We evaluate the effectiveness of canonical points refined with F_s and Conv-Filter.

	PSNR↑	SSIM↑	†LPIPS↓
w/o F_s refine	30.93	0.971	24.712
w/o filter	9.22	0.607	318.460
Full	**31.09**	**0.974**	**24.085**

5. Limitations

Our method is state-of-the-art in the task of learning implicit human representations from limited input and synthesizing diverse pose images. What's required is only a monocular video, even a few images, and easily obtainable SMPL information, without the need for additional calculations of texture information, greatly expanding the method's universality. However, our method still has certain limitations: 1) The effectiveness of our method depends on the accuracy of the estimated SMPL, and when the SMPL accuracy is low, the results may be blurred. Currently, the accuracy of SMPL estimation methods is not always satisfactory. 2) Our method uses coordinate voxelization to assist calculation, which may cause image edge serrations. Fine-tuning can be achieved by adjusting the voxel size and convolution kernel size, which will increase computational cost. 3) Our method uses basic SMPL information for training, so it is difficult to drive hand and facial details.

6. Conclusion

We propose a human neural radiance field model that can train with limited inputs and generalize to diverse zero-shot poses. Unlike previous methods, our approach filters the sampling points and obtains point-level features in a voxel volume of explicit vertices, and subsequently deforms the points to a canonical space using general and specific mapping. Our approach uses less than 1% learnable parameters and achieves state-of-the-art novel pose metrics in our experiments, and maintains the best performance with few-shot input. Our approach only requires estimated SMPL information, which can be easily obtained using existing methods, thereby maintaining usability while being able to generalize to industries such as video production.

Acknowledgements. This work was supported by the Hubei Key R&D Project (2022BAA033), the National Natural Science Foundation of China (62171325), the Beijing Nova Program (20220484063), and the Value Exchange Engineering, a joint research project between Mercari, Inc. and RIISE. The numerical calculations in this paper have been done on the supercomputing system in the Supercomputing Center of Wuhan University.

References

[1] Kara-Ali Aliev, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov, and Victor Lempitsky. Neural point-based graphics. In *Proceedings of the European Conference on Computer Vision*, 2020. 3

[2] 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. 3

[3] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3D generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3

[4] Mingfei Chen, Jianfeng Zhang, Xiangyu Xu, Lijuan Liu, Yujun Cai, Jiashi Feng, and Shuicheng Yan. Geometry-guided progressive NeRF for generalizable and efficient neural human rendering. In *Proceedings of the European Conference on Computer Vision*, 2022. 3

[5] Yue Chen, Xuan Wang, Xingyu Chen, Qi Zhang, Xiaoyu Li, Yu Guo, Jue Wang, and Fei Wang. Uv volumes for real-time rendering of editable free-view human performance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

[6] Wei Cheng, Su Xu, Jingtan Piao, Chen Qian, Wayne Wu, Kwan-Yee Lin, and Hongsheng Li. Generalizable neural performer: Learning robust radiance fields for human novel view synthesis. *arXiv preprint arXiv:2204.11798*, 2022. 3

[7] SpConv Contributors. Spconv: Spatially sparse convolution library. `https://github.com/traveller59/spconv`, 2022. 4

[8] John Flynn, Michael Broxton, Paul Debevec, Matthew DuVall, Graham Fyffe, Ryan Overbeck, Noah Snavely, and Richard Tucker. Deepview: View synthesis with learned gradient descent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 3

[9] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 3

[10] Xiangjun Gao, Jiaolong Yang, Jongyoo Kim, Sida Peng, Zicheng Liu, and Xin Tong. Mps-NeRF: Generalizable 3D human rendering from multiview images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[11] Chen Geng, Sida Peng, Zhen Xu, Hujun Bao, and Xiaowei Zhou. Learning neural volumetric representations of dynamic humans in minutes. *arXiv preprint arXiv:2302.12237*, 2023. 1, 2, 3

[12] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 3

[13] 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*, 2021. 3

[14] Vinoj Jayasundara, Amit Agrawal, Nicolas Heron, Abhinav Shrivastava, and Larry S Davis. FlexNeRF: Photorealistic free-viewpoint rendering of moving humans from sparse views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

[15] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3

[16] Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Instantavatar: Learning avatars from monocular video in 60 seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2, 3

[17] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *Proceedings of the European Conference on Computer Vision*, 2022. 1, 2, 3, 5

[18] 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*, 34:24741–24752, 2021. 1, 3

[19] Guopeng Li, Ming Qian, and Gui-Song Xia. Unleashing unlabeled data: A paradigm for cross-view geo-localization. *arXiv preprint arXiv:2403.14198*, 2024. 3

[20] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhöfer, Jürgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. In *Proceedings of the European Conference on Computer Vision*, 2022. 3

[21] 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*, 2021. 3

[22] Zhe Li, Zerong Zheng, Yuxiao Liu, Boyao Zhou, and Yebin Liu. Posevocab: Learning joint-structured pose embeddings for human avatar modeling. *arXiv preprint arXiv:2304.13006*, 2023. 3

[23] Yiyi Liao, Katja Schwarz, Lars Mescheder, and Andreas Geiger. Towards unsupervised learning of generative models for 3D controllable image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020. 3

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

[25] 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. 3

[26] 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 Transactions on Graphics (TOG)*, 40(6):1–16, 2021. 1, 2, 3

[27] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 3

[28] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751*, 2019. 3

[29] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 2, 3

[30] Ricardo Martin-Brualla, Rohit Pandey, Shuoran Yang, Pavel Pidlypenskyi, Jonathan Taylor, Julien Valentin, Sameh Khamis, Philip Davidson, Anastasia Tkach, Peter Lincoln, et al. Lookingood: Enhancing performance capture with real-time neural re-rendering. *arXiv preprint arXiv:1811.05029*, 2018. 3

[31] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 3

[32] Nelson Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1(2):99–108, 1995. 5

[33] 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 *Proceedings of the European Conference on Computer Vision*, 2020. 1, 3

[34] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 3

[35] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 3

[36] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3D representations without 3D supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 3

[37] 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*, 2021. 1, 3

[38] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. HyperNeRF: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021. 1, 3

[39] Bo Peng, Jun Hu, Jingtao Zhou, Xuan Gao, and Juyong Zhang. Intrinsicngp: Intrinsic coordinate based hash encoding for human NeRF. *arXiv preprint arXiv:2302.14683*, 2023. 3

[40] 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*, 2021. 1, 2, 3, 5, 6, 7

[41] 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*, 2021. 1, 3, 4, 5

[42] 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*, 2021. 1, 3

[43] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, 2019. 3

[44] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 3

[45] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3D feature embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 3

[46] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3D-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[47] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 3

[48] 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*, 34:12278–12291, 2021. 3

[49] Shih-Yang Su, Timur Bagautdinov, and Helge Rhodin. Npc: Neural point characters from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 3

[50] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. Monocular, one-stage, regression of multiple 3D people. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021. 2

[51] 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, 2020. 3

[52] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 3

[53] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 1, 3

[54] Chen Wang, Xian Wu, Yuan-Chen Guo, Song-Hai Zhang, Yu-Wing Tai, and Shi-Min Hu. NeRF-sr: High quality neural radiance fields using supersampling. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022. 1, 3

[55] 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*, 2021. 1, 3

[56] 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*, 2022. 1, 2, 3, 5, 6, 7, 4

[57] Minye Wu, Yuehao Wang, Qiang Hu, and Jingyi Yu. Multi-view neural human rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 3

[58] Zijin Wu, Xingyi Li, Juewen Peng, Hao Lu, Zhiguo Cao, and Weicai Zhong. Dof-NeRF: Depth-of-field meets neural radiance fields. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022. 1, 3

[59] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 1, 3

[60] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J Black. Icon: Implicit clothed humans obtained from normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3

[61] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. Monoperfcap: Human performance capture from monocular video. *ACM Transactions on Graphics (ToG)*, 37 (2):1–15, 2018. 3

[62] Zhengming Yu, Wei Cheng, Xian Liu, Wayne Wu, and Kwan-Yee Lin. Monohuman: Animatable human neural field from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1, 2, 3, 5, 6, 7, 4

[63] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. NeRF++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 1, 3

[64] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 5

[65] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. NeRFactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics (TOG)*, 40(6):1–18, 2021. 3

[66] Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. HumanNeRF: Efficiently generated human radiance field from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 3

[67] Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. Structured local radiance fields for human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3

[68] Zerong Zheng, Xiaochen Zhao, Hongwen Zhang, Boning Liu, and Yebin Liu. Avatarrex: Real-time expressive full-body avatars. *arXiv preprint arXiv:2305.04789*, 2023. 3

[69] Shihao Zhou, Mengxi Jiang, Shanshan Cai, and Yunqi Lei. Dc-gnet: Deep mesh relation capturing graph convolution network for 3D human shape reconstruction. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021. 3

[70] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 3