This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

HumanNeRF:
Free-viewpoint Rendering of Moving People from Monocular Video

Chung-Yi Weng[1] Brian Curless[1,2] Pratul P. Srinivasan[2] Jonathan T. Barron[2] Ira Kemelmacher-Shlizerman[1,2]
[1]University of Washington [2]Google Research

(a) monocular video (b) free-viewpoint rendering

Figure 1. Our method takes as input a monocular video[1] of a human performing complex movement, e.g., dancing (left), and creates a free-viewpoint rendering for any frame in the sequence (right). We construct a canonical subject appearance volume, and a motion field mapping from observation to canonical space, trained on the video. At test time, we take just the pose from the source frame (red square) and synthesize all output views, including the target view. Please refer to the project page[2] to see the animated results.

Abstract

We introduce a free-viewpoint rendering method – HumanNeRF – that works on a given monocular video of a human performing complex body motions, e.g. a video from YouTube. Our method enables pausing the video at any frame and rendering the subject from arbitrary new camera viewpoints or even a full 360-degree camera path for that particular frame and body pose. This task is particularly challenging, as it requires synthesizing photorealistic details of the body, as seen from various camera angles that may not exist in the input video, as well as synthesizing fine details such as cloth folds and facial appearance. Our method optimizes for a volumetric representation of the person in a canonical T-pose, in concert with a motion field that maps the estimated canonical representation to every frame of the video via backward warps. The motion field is decomposed into skeletal rigid and non-rigid motions, produced by deep networks. We show significant performance improvements over prior work, and compelling examples of free-viewpoint renderings from monocular video of moving humans in challenging uncontrolled capture scenarios.

1. Introduction

Given a single video of a human performing an activity, e.g., a YouTube or TikTok video of a dancer, we would like the ability to pause at any frame and rotate 360 degrees around the performer to view them from any angle at that moment in time (Figure 1). This problem – free-viwepoint rendering of a moving subject – is a longstanding research challenge, as it involves synthesizing previously unseen camera views while accounting for cloth folds, hair movement, and complex body poses [4, 5, 14, 17, 26, 37, 57, 63]. The problem is particularly hard for the case of "in-the-wild" videos taken with a single camera (monocular video), the case we address in this paper.

Previous neural rendering methods [2, 32, 35, 36, 48, 64, 73] typically assume multi-view input, careful lab capture, or do not perform well on humans due to non-rigid body motion. Human-specific methods typically assume a SMPL template [33] as a prior, which helps constrain the motion space but also introduces artifacts in clothing and complex motions that are not captured by the SMPL model [47, 48]. Recently deformable NeRF methods [45, 46, 49, 62] perform well for small deformations, but not for large, full body motions like dancing.

We introduce a method, called HumanNeRF, that takes as input a single video of a moving person and, after per-frame, off-the-shelf segmentation (with some manual clean-

---

[1] e.g., `https://youtu.be/0ORaAnJYROg`
[2] `https://grail.cs.washington.edu/projects/humannerf/`

up) and automatic 3D pose estimation, optimizes for a canonical, volumetric T-pose of the human together with motion field that maps the estimated canonical volume to each video frame via a backward warping. The motion field combines skeletal rigid motion with non-rigid motion, each represented volumetrically. Our solution is data-driven, with the canonical volume and motion fields derived from the video itself and optimized for large body deformations, trained end-to-end, including 3D pose refinement, without template models. At test time, we can pause at any frame in the video and, conditioned on the pose in that frame, render the resulting volumetric representation from any viewpoint.

We show results on a variety of examples: existing lab datasets, videos we captured outside the lab, and downloads from YouTube (with creator permission). Our method outperforms the state-of-the-art numerically and produces significantly higher visual quality. Please refer to the project page to see the results in motion.

2. Related Work

The physics of free-viewpoint rendering involves modeling geometry and surface properties and then rendering from new camera views. However, it remains difficult to recreate complex geometry and subtle lighting effects. Alternatively, image-based rendering [55, 59] offers to render novel views based on given set of views in the image domain with a large corpus of research over the last couple decades [7, 8, 12, 16, 20, 21, 29, 76].

Human specific rendering: The work of Kanade et al. [26] is one of the earliest investigations into free-viewpoint rendering of humans. It introduced a dome equipped with cameras to recover depth maps and meshes, enabling novel views to be rendered by reprojecting and blending different views to account for mesh holes due to occlusions. Later, Matusik et al. [37] reconstructed a *visual hull* from silhouettes of the subject and rendered it by carefully selecting pixels without an auxiliary geometric representation. Carranza et al. [4] used a parameterized body model as a prior and combined marker-less motion capture and view-dependent texturing [12]. Follow-on work introduced non-rigid deformation [63], texture warping [5, 70], and various representations based on volumes [11] or spheres [57]. Collet et al. [10] and Guo et al. [17] build a system as well as pipeline that produces high-quality streamable [10] or even relightable [17] free-viewpoint videos of moving people.

Most of these methods rely on multi-view videos – typically expensive studio setups – while we are interested in a simple monocular camera configuration.

Neural radiance fields: NeRF [40] and its extensions [2, 22, 42, 56, 60, 73, 75] enable high quality rendering of novel views of static scenes. NeRF has recently been extended to dynamic scenes [15, 30, 45, 46, 49, 62, 69], though these approaches generally assume that motion is small. We compare our method to these dynamic and deformable NeRF works in our results section.

Human-specific neural rendering: The work of Liu et al. [32] starts from a pre-captured body model and learns to model time-dependent dynamic textures and enforce temporal coherence. Martin-Brualla et al. [35] trained a UNet to improve the artifacts introduced by volumetric capture. The follow-up work of Pandey et al. [44] reduced the number of required input frames to as few as a single RGBD image via semi-parametric learning. Wu et al. [68] and Peng et al. [48] explored the use of learned structured latent codes embedded for point clouds (from MVS [53]) or reposed mesh vertices (from SMPL [33]) and learn an accompanying UNet- or NeRF-based neural renderer. Zhang et al. [25] decomposed a scene into background and individual performers, and represented them with separated NeRFs thus enabling scene editing. Other than free-viewpoint rendering, there is another related active research field that focus on human motion retargeting either in 2D [1, 6, 34, 41, 52, 65, 66] or 3D [18, 19, 24, 31, 47, 51, 67, 72]. The main difference between our method and those works is that we take as input *monocular* video that contains *complex* human motions and enable high-fidelity full 3D rendering.

Additionally, our formulation of skeletal motion draws inspiration from Vid2Actor proposed by Weng et al. [67], a method intended for rigidly animatable characters. Instead, we focus on the free-viewpoint application and recovering pose-dependent, non-rigid deformation and outperform them significantly for this application.

Concurrent work: Xu et al. [71] co-learn implicit geometry as well as appearance from images. They largely focus on multi-view setups with a few examples on monocular videos where the human motion is simple (A-pose). Su et al. [58] use an over-parameterized NeRF to rigidly transform NeRF features for refining body pose and thus final rendering. The non-rigid motion is not explicitly modeled and the rendering quality is not high. A similar approach is discovered by Noguchi et al. [43] as well but still shows results of limited visual quality.

3. Representing a Human as a Neural Field

We represent a moving person with a canonical appearance volume F_c warped to an observed pose to produce output appearance volume F_o:

$$F_o(\mathbf{x}, \mathbf{p}) = F_c(T(\mathbf{x}, \mathbf{p})), \tag{1}$$

where $F_c : \mathbf{x} \rightarrow (\mathbf{c}, \sigma)$ maps position $\mathbf{x}$ to color $\mathbf{c}$ and density σ, and $T : (\mathbf{x}_o, \mathbf{p}) \rightarrow \mathbf{x}_c$ defines a motion field mapping points from observed space back to canonical space, guided by observed pose $\mathbf{p} = (J, \Omega)$, where J includes K standard 3D joint locations, and $\Omega = \{\boldsymbol{\omega}_i\}$ are local joint rotations represented as axis-angle vectors $\boldsymbol{\omega}_i$.

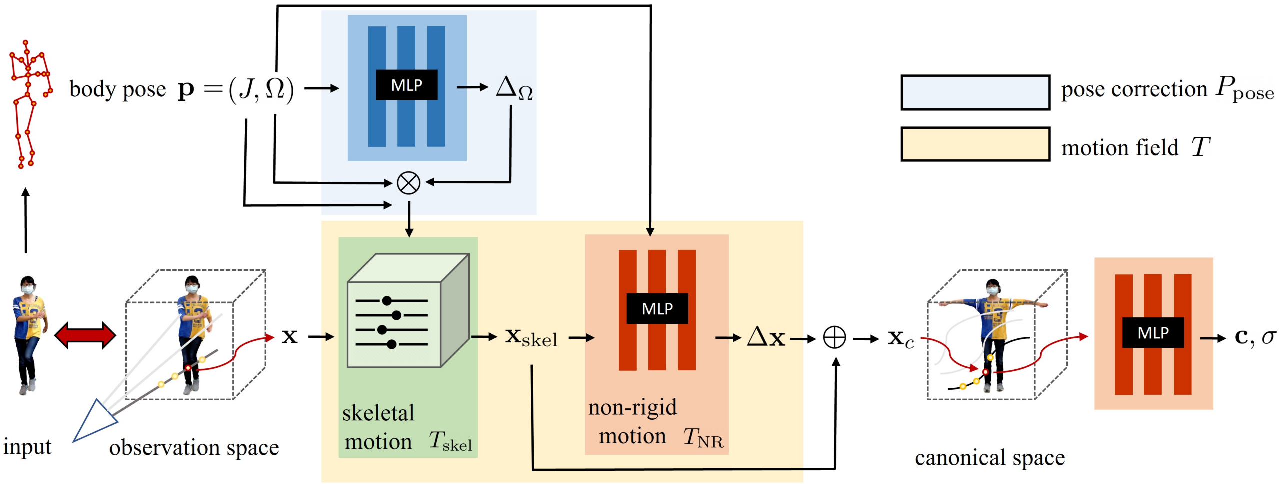


Figure 2. Our method takes a video frame as input and optimizes for canonical appearance, represented as a continuous field, as well as a motion field mapping from observation to canonical space. The motion field is decomposed into skeletal rigid and non-rigid motion, represented as a discrete grid and a continuous field respectively. We additionally refine body pose initialized with an off-the-shelf body pose estimator, leading to better alignment. A loss is imposed between the volume rendering in observation space and the input image, directing optimization towards a solution.

We handle complex human movement with complex deformation by decomposing the motion field into two parts:

$$T(\mathbf{x}, \mathbf{p}) = T_{\text{skel}}(\mathbf{x}, \mathbf{p}) + T_{\text{NR}}(T_{\text{skel}}(\mathbf{x}, \mathbf{p}), \mathbf{p}), \tag{2}$$

where T_{skel} represents skeleton-driven deformation, essentially inverse (volumetric) linear-blend skinning, and T_{NR} starts from the skeleton-driven deformation and produces an offset $\Delta\mathbf{x}$ to it. In effect, T_{skel} provides the coarse deformation driven by standard skinning, and T_{NR} provides the more non-rigid effects, e.g., due to deformation of clothing.

For "in-the-wild" imagery, we use an off-the-shelf 3D body+camera pose estimator. Due to inaccuracy in pose estimation, we also solve for a pose correction function $P_{\text{pose}}(\mathbf{p})$ that better explains the observations, and apply this improvement to the skeleton-driven deformation, i.e., we replace $T_{\text{skel}}(\mathbf{x}, \mathbf{p})$ with $T_{\text{skel}}(\mathbf{x}, P_{\text{pose}}(\mathbf{p}))$ in Eq. 2.

Figure 2 gives an overview of the components of our system. In the following sections, we describe these components in detail.

Canonical volume: We represent the canonical volume F_c as a continuous field with an MLP that outputs color $\mathbf{c}$ and density σ given a point $\mathbf{x}$:

$$F_c(\mathbf{x}) = \text{MLP}_{\theta_c}(\gamma(\mathbf{x})), \tag{3}$$

where γ is a sinusoidal positional encoding defined as $(\mathbf{x}, \sin(2^0\pi\mathbf{x}), \cos(2^0\pi\mathbf{x}), ..., \sin(2^{L-1}\pi\mathbf{x}), \cos(2^{L-1}\pi\mathbf{x}))$ and L is a hyper-parameter that determines the number of frequency bands [40].

Skeletal motion: Following Weng et al. [67], we compute the skeletal deformation T_{skel} as a kind of inverse, linear blend skinning that maps points in the observation space to the canonical space:

$$T_{\text{skel}}(\mathbf{x}, \mathbf{p}) = \sum_{i=1}^{K} w_o^i(\mathbf{x})(R_i\mathbf{x} + \mathbf{t}_i), \tag{4}$$

where w_o^i is the blend weight for the i-th bone and R_i, $\mathbf{t}_i$ are the rotation and translation, respectively, that map the bone's coordinates from observation to canonical space; R_i and $\mathbf{t}_i$ can be explicitly computed from $\mathbf{p}$ (see supplementary). We then aim to optimize for w_o^i.

In practice, we solve for w_c^i defined in canonical space by storing K blend weights as a set of volumes $\{w_c^i(\mathbf{x})\}$, from which the observation weights are derived as:

$$w_o^i(\mathbf{x}) = \frac{w_c^i(R_i\mathbf{x} + \mathbf{t}_i)}{\sum_{k=1}^{K} w_c^k(R_k\mathbf{x} + \mathbf{t}_k)}. \tag{5}$$

Solving for a single set of weight volumes $\{w_c^i(\mathbf{x})\}$ in canonical space, instead of N sets of $\{w_o^i(\mathbf{x})\}$ in observation space (corresponding to N input images), can lead to better generalization as it avoids over-fitting [9, 67].

We pack the set of $\{w_c^i(\mathbf{x})\}$ into a single volume $W_c(\mathbf{x})$ with K channels. Rather than encode W_c with an MLP, we choose an explicit volume representation for two reasons: (1) Eq. 5 shows that K MLP evaluations would be needed to compute each $w_o^i(\mathbf{x})$, infeasible for optimization ($K = 24$ in our work); (2) an explicit volume with limited resolution resampled via trilinear interpolation provides smoothness that can help regularize the optimization later. In practice, during optimization, rather than directly solve for volume W_c, we solve for parameters θ_{skel} of a CNN that generates the volume from a random (constant) latent code $\mathbf{z}$:

$$W_c(\mathbf{x}) = \text{CNN}_{\theta_{\text{skel}}}(\mathbf{x}; \mathbf{z}). \tag{6}$$

We also add one more channel, a background class, and represent W_c as a volume with $K+1$ channels. We then apply channel-wise *softmax* to the output of the CNN, enforcing a partition of unity across the channels. The denominator of Eq. 5 can then be used to approximate likelihood $f(\mathbf{x})$ of being part of the subject, where $f(\mathbf{x}) = \sum_{k=1}^{K} w_c^k(R_k\mathbf{x}+\mathbf{t}_k)$. When $f(\mathbf{x})$ is close to zero, we are likely in free space away from the subject, which we will use during volume rendering.

The idea of optimizing blend weights (or skinning field) is not new. Similar approaches have been applied to human modeling [3, 9, 13, 24, 39, 47, 50, 61, 72]. Our formulation follows Weng et al. [67], but also shares similarities with Tiwari et al. [61]; the latter learns from 3D scans while we learn from 2D images, like the former.

Non-rigid motion: We represent non-rigid motion T_{NR} as an offset $\Delta\mathbf{x}$ to the skeleton-driven motion, conditioned on that motion, i.e., $\Delta\mathbf{x}(\mathbf{x},\mathbf{p}) = T_{\text{NR}}(T_{\text{skel}}(\mathbf{x},\mathbf{p}),\mathbf{p}))$. To capture detail, we represent T_{NR} with an MLP:

$$T_{\text{NR}}(\mathbf{x},\mathbf{p}) = \text{MLP}_{\theta_{\text{NR}}}(\gamma(\mathbf{x});\Omega), \tag{7}$$

where again we use the standard positional encoding γ and condition the MLP on Ω, the joint angles of body pose $\mathbf{p}$.

Pose correction: The body pose $\mathbf{p} = (J,\Omega)$ estimated from an image is often inaccurate. To address this, we solve for an update to the pose:

$$P_{\text{pose}}(\mathbf{p}) = (J, \Delta_\Omega(\mathbf{p}) \otimes \Omega), \tag{8}$$

where we hold the joints J fixed and optimize for a relative update to the joint angles, $\Delta_\Omega = (\Delta\boldsymbol{\omega}_0, ..., \Delta\boldsymbol{\omega}_K)$ which is then applied to Ω to get updated rotation vectors.

Empirically we found, instead of directly optimizing for Δ_Ω, solving for the parameters θ_{pose} of an MLP that generates Δ_Ω conditioned on Ω leads to faster convergence:

$$\Delta_\Omega(\mathbf{p}) = \text{MLP}_{\theta_{\text{pose}}}(\Omega). \tag{9}$$

With this pose correction, we can re-write the equation that warps from observation space to canonical space as:

$$T(\mathbf{x},\mathbf{p}) = T_{\text{skel}}(\mathbf{x}, P_{\text{pose}}(\mathbf{p})) + T_{\text{NR}}(T_{\text{skel}}(\mathbf{x}, P_{\text{pose}}(\mathbf{p})), \mathbf{p}) \tag{10}$$

4. Optimizing a HumanNeRF

In this section, we describe the overall objective function we minimize, our volume rendering procedure, how we regularize the optimization process, specific loss function details, and the ray sampling method.

HumanNeRF objective: Given input frames $\{I_1, I_2, ..., I_N\}$, body poses $\{\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_N\}$, and cameras $\{\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_N\}$, we are solving the problem:

$$\underset{\Theta}{\text{minimize}} \quad \sum_{i=1}^{N} \mathcal{L}\{\Gamma[F_c(T(\mathbf{x},\mathbf{p}_i)), \mathbf{e}_i], I_i\}, \tag{11}$$

where $\mathcal{L}\{\cdot\}$ is the loss function and $\Gamma[\cdot]$ is a volume renderer, and we minimize the loss with respect to all network parameters $\Theta = \{\theta_c, \theta_{\text{skel}}, \theta_{\text{NR}}, \theta_{\text{pose}}\}$. As we have seen, F_c is determined by parameters θ_c, while the transformation T from observation space to canonical space relies on parameters θ_{skel}, θ_{NR}, and θ_{pose}.

4.1. Volume rendering

We render a neural field using the volume rendering equation [38] as described by Mildenhall et al. [40]. The expected color $C(\mathbf{r})$ of a ray $\mathbf{r}$ with D samples can be written as:

$$\begin{aligned} C(\mathbf{r}) &= \sum_{i=1}^{D}\Big(\prod_{j=1}^{i-1}(1-\alpha_j)\Big)\alpha_i \mathbf{c}(\mathbf{x}_i), \\ \alpha_i &= 1 - \exp(-\sigma(\mathbf{x}_i)\Delta t_i), \end{aligned} \tag{12}$$

where Δt_i is the interval between sample i and $i+1$.

We further augment the definition of α_i to be small when approximate foreground probability $f(\mathbf{x})$ is low:

$$\alpha_i = f(\mathbf{x}_i)(1 - \exp(-\sigma(\mathbf{x}_i)\Delta t_i)), \tag{13}$$

We apply the stratified sampling approach proposed by NeRF [40]. We do not use hierarchical sampling since the bounding box of a subject can be estimated from their 3D body pose. We then only sample points inside the box.

4.2. Delayed optimization of non-rigid motion field

When solving for all the network parameters in Eq. 11 at once, we find that the the optimized skeleton-driven and non-rigid motions are not decoupled – a portion of the subject's skeletal motions is modeled by the non-rigid motion field – due to over-fitting of non-rigid motions to the input images. As a result, the quality degrades when rendering unseen views.

We manage the optimization process to solve the problem. Specifically, we disable non-rigid motions at the beginning of optimization, and then bring them back in a coarse-to-fine manner [23, 45]. To achieve this, for the non-rigid motion MLP, we apply a truncated Hann window to its frequency bands of positional encoding, to prevent overfitting to the data [60], increasing the window size as the optimization proceeds. Following Park et al. [45], we define the weight for each frequency band j of positional encoding:

$$w(\tau) = \frac{1 - \cos(\text{clamp}(\tau - j, 0, 1)\pi)}{2}, \tag{14}$$

where $\tau \in [0, L)$ determines the width of a truncated Hann window, and L is the total number of frequency bands in positional encoding. We then define τ as a function of the optimization iteration:

$$\tau(t) = L\frac{\max(0, t - T_s)}{T_e - T_s}, \tag{15}$$

	Subject **377**			Subject **386**			Subject **387**		
	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓
Neural Body [48]	29.11	0.9674	40.95	30.54	0.9678	46.43	27.00	0.9518	59.47
Ours	30.41	0.9743	24.06	33.20	0.9752	28.99	28.18	0.9632	35.58

	Subject **392**			Subject **393**			Subject **394**		
	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓	PSNR ↑	SSIM ↑	LPIPS* ↓
Neural Body [48]	30.10	0.9642	53.27	28.61	0.9590	59.05	29.10	0.9593	54.55
Ours	31.04	0.9705	32.12	28.31	0.9603	36.72	30.31	0.9642	32.89

Table 1. Quantitative comparison on ZJU-MoCap dataset. We color cells that have the best metric value. LPIPS* = LPIPS $\times 10^3$.

where t is the current iteration, and T_s and T_e are hyperparameters that determine when to enable non-rigid motion optimization and when to use full frequency bands of positional encoding. We remove position identity from positional encoding without affecting performance [2]. By doing so, we can completely disable non-rigid motion optimization by setting $\tau = 0$ [46].

4.3. Loss and ray sampling

Loss function: We employ both an MSE loss to match pixel-wise appearance and a perceptual loss, LPIPS [74], to provide robustness to slight misalignments and shading variation and to improve detail in the reconstruction. Our final loss function is $\mathcal{L} = \mathcal{L}_{\text{LPIPS}} + \lambda \mathcal{L}_{\text{MSE}}$. We use $\lambda = 0.2$ and choose VGG as the backbone of LPIPS.

Patch-based ray sampling: Training on random ray samples, as done in NeRF [40], does not work for minimizing our loss because LPIPS uses convolutions to extract features. Instead, we sample G patches with size $H \times H$ on an image, and render a total of $G \times H \times H$ rays in each batch. The rendered patch is compared against the patch with the same position on the input image. We use $G = 6$ and $H = 32$ in our experiments. Similar approaches were also used in NeRF-based generative models [54].

5. Results

5.1. Evaluation dataset

We evaluate our method on the ZJU-MoCap dataset [48], self-captured data (*rugby, hoodie*), and YouTube videos downloaded from Internet (*story*[3], *way2sexy*[4], *invisible*[5]). All subjects in these videos provided consent to use their data. For ZJU-MoCap, we select 6 subjects (377, 386, 387, 392, 393, 394) with diverse motions and use images captured by "*camera 1*" as input and the other 22 cameras for evaluation. We directly apply camera matrices, body pose, and segmentation provided by the dataset. For videos "in the wild" (self-captured and YouTube videos), we run SPIN [28] to get approximate camera and body pose, automatically segment the foreground subject, and then manually correct errors in the segmentation. (High quality segmentation is necessary for best results; purely automatic segmenters were not accurate enough, and improving on them was outside the scope of this paper, an area of future work.) We additionally resize video frames to keep the height of subject at approximately 500 pixels.

5.2. Optimization details

We optimize Eq. 11 using the Adam optimizer [27] with hyperparameters $\beta_1 = 0.9$ and $\beta_2 = 0.99$. We set the learning rate to 5×10^{-4} for θ_c (the canonical MLP), and 5×10^{-5} for all the others. We use 128 samples per ray. The optimization takes $400K$ iterations (about 72 hours) on 4 GeForce RTX 2080 Ti GPUs. We apply delayed optimization with $T_s = 10K$ and $T_e = 50K$ to ZJU-MoCap data, and with $T_s = 100K$ and $T_e = 200K$ to the others. In addition, we postpone pose refinement until after 20K iterations for in-the-wild videos.

5.3. Evaluation method

	Neural Body	HyperNeRF	HumanNeRF
Setup	multi-camera	single camera	single camera
Subject	dynamic human	quasi-static general scene	dynamic human
Priors	body pose, SMPL vertices (reposed)	rigidity	body pose (approx.)

Table 2. Differences between the compared methods.

We compare our method with Neural Body [48] (typically used with multiple cameras) and HyperNeRF [46] (single moving camera around the subject), state-of-the-art methods for modeling humans and general scenes for novel view synthesis. Our method works with a single camera which can be static or moving; we focus on results with a static camera and moving subjects, a natural way to capture a person's performance. The differences between these methods are listed in Table 2.

---

[3] `https://youtu.be/0ORaAnJYROg`
[4] `https://youtu.be/gEpJDE8ZbhU`
[5] `https://youtu.be/ANwEiICt7BM`

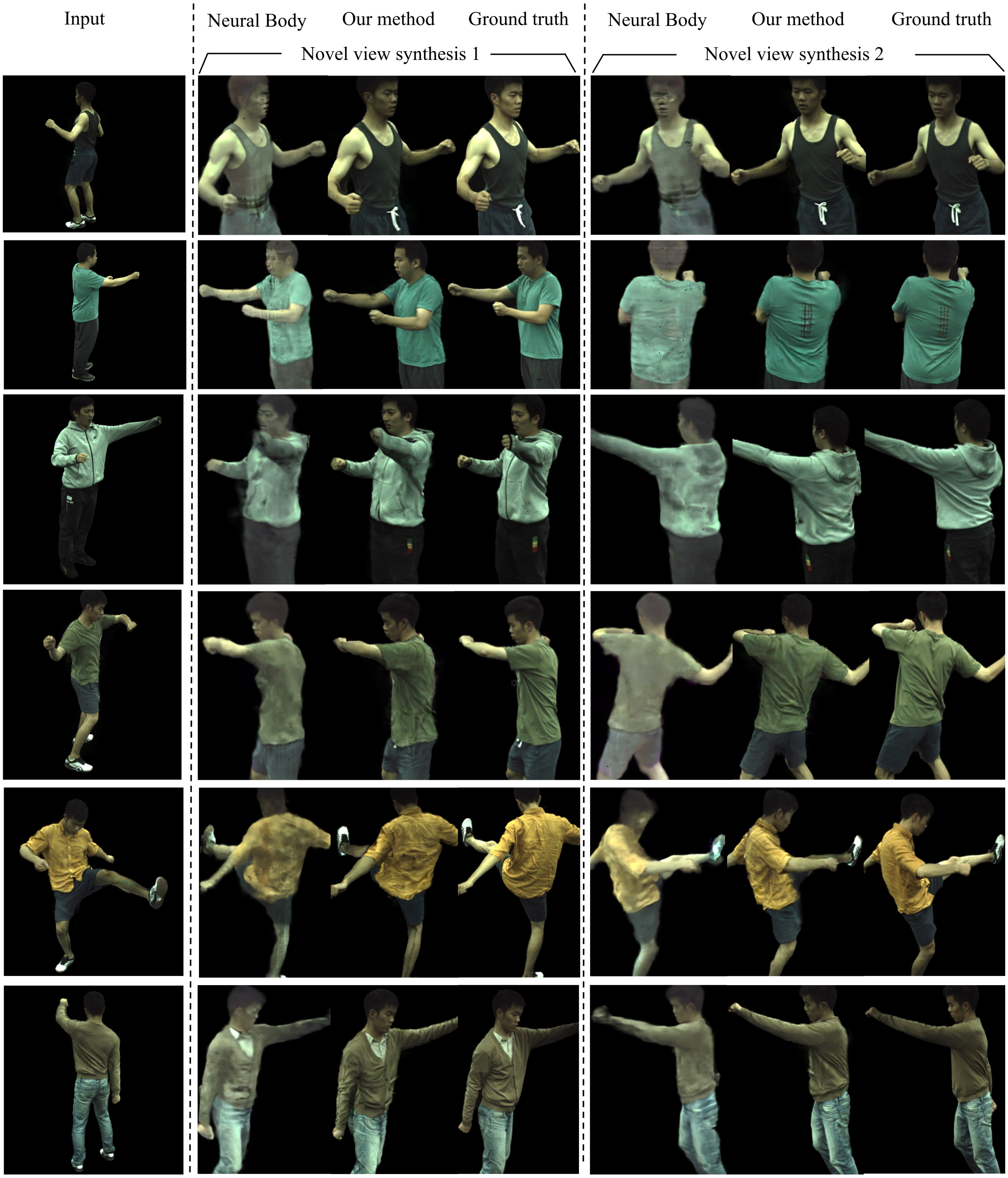
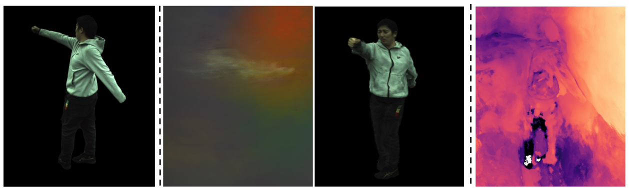


Figure 3. Qualitative comparison on ZJU-MoCap dataset.

5.4. Comparisons

We found HyperNeRF does not produce meaningful output for novel view synthesis in our experiments, as shown in Fig. 4, likely because it relies on multiple views (moving camera) to build a coherent 3D model. For the static camera case with moving subject, it fails to recover a meaningful depth map and appears to memorizes the input images rather than generalize from them. We note that dynamic



(a) Input (b) HyperNeRF (c) Our method (d) Learned depth (HyperNeRF)

Figure 4. Qualitative comparison to HyperNeRF [46]

human motions are also more extreme than the examples shown to work with HyperNeRF.

Quantitatively, as shown in Table 1, HumanNeRF outperforms Neural Body for all subjects and under all metrics, except for subject 393 on PSNR (a metric known to favor smooth results [74]). The gain is particularly significant with perceptual metric LPIPS, nearly 40% improvement on average. Fig. 3 shows that HumanNeRF's visual quality is substantially better then Neural Body for this dataset. Our method is capable of producing high fidelity details similar to the ground truth even on completely unobserved views, while Neural Body tends to produce blurrier results. The results for self-captured and YouTube videos, shown in Fig. 5, also show consistently higher quality reconstructions with HumanNeRF.

5.5. Ablation studies

(a) input (b) w/o pose correction w/o non-rigid motion (c) w/o non-rigid motion (d) full model

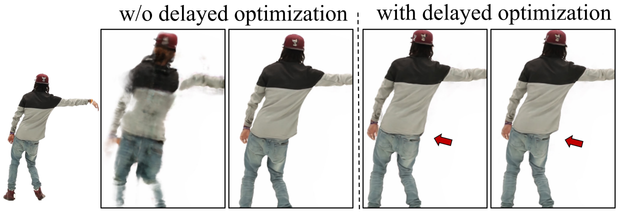
Figure 6. Pose correction and non-rigid motion improve novel view synthesis. Pose correction straightens the right arm and adds details (red arrows in (b) vs (c)) and non-rigid deformation improves clothing alignment and shape (green arrows in (c) vs. (d)).

Table 3 illustrates that skeletal deformation alone is enough for significant improvement over Neural Body for the ZJU-MoCap data. Adding non-rigid deformation provides further gains. (Accurate poses were provided for this dataset, thus we did not perform an ablation for the pose optimizer here.)

Fig. 6 shows visually, for in-the-wild data, the importance of including non-rigid motion and, additionally, pose correction for an unseen view.

	PSNR ↑	SSIM ↑	LPIPS* ↓
Neural Body [48]	29.08	0.9616	52.27
Ours (w/o non-rigid)	29.81	0.9657	34.17
Ours (full model)	30.24	0.9679	31.73

Table 3. Ablation study on ZJU-MoCap. We compute averages over 6 sequences. We color cells with best best and second best metric values. LPIPS* = LPIPS $\times 10^3$.



(a) input (b) skeletal only (c) full motion (d) skeletal only (e) full motion

Figure 7. Delayed optimization (d, e) leads to better motion decoupling than the result without it (b, c). The skeletal-only deformation result without delayed optimization is poor, which can "corrected" by the non-rigid deformation, but leads to poor view generalization (below).

Fig. 7 shows the importance of delayed optimization for decoupling skeletal deformation and non-rigid deformation. When not decoupled well, generalization to new views is much poorer, as shown in Fig. 8.

(a) input (b) w/o delayed optimization (c) with delayed optimization

Figure 8. Without delayed optimization and strong decoupling of skeletal and non-rigd deformations, generalization to unseen views is poor (b). With delayed optimization, the decoupling leads to good generalization (c).

6. Discussion

Limitations: Our method has artifacts when part of the body is not shown in the video. Pose correction improves image alignment but may fail if the initial pose estimate is poor or if the image contains strong artifacts such as motion blur. In addition, we observed the frame-by-frame body poses are still not temporally smooth even after pose correction. We assume non-rigid motion is pose-dependent, but this is not always true (e.g., clothes shifting due to wind or due to follow-through after dynamic subject motion). We also assume fairly diffuse lighting, so that appearance does not change dramatically as the points on the subject rotate

Figure 5. Qualitative comparison for self-captured videos (first two rows) and YouTube videos (bottom three).

around. Finally, for in-the-wild videos, we rely on manual intervention to correct segmentation errors. These limitations point to a range of interesting avenues for future work.

Conclusion: We have presented HumanNeRF, producing state-of-the-art results for free-viewpoint renderings of moving people from monocular video. We demonstrate high fidelity results for this challenging scenario by carefully modeling body pose and motion as well as regularizing the optimization process. We hope the result points in a promising direction toward modeling humans in motion and, eventually, achieving fully photorealistic, free-viewpoint rendering of people from casual captures.

Acknowledgement: We thank Marquese Scott for generously allowing us to feature his inspiring videos in this work. Special thanks to dear Lulu Chu for her enduring support. This work was funded by the UW Reality Lab, Meta, Google, Futurewei, and Amazon.

References

[1] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. *CVPR*, 2018. 2

[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. *ICCV*, 2021. 1, 2, 5

[3] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration. *Advances in Neural Information Processing Systems*, 33, 2020. 4

[4] Joel Carranza, Christian Theobalt, Marcus A Magnor, and Hans-Peter Seidel. Free-viewpoint video of human actors. *ACM transactions on graphics (TOG)*, 2003. 1, 2

[5] Dan Casas, Marco Volino, John Collomosse, and Adrian Hilton. 4D video textures for interactive character appearance. *Computer Graphics Forum*, 2014. 1, 2

[6] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. *ICCV*, 2019. 2

[7] Gaurav Chaurasia, Sylvain Duchene, Olga Sorkine-Hornung, and George Drettakis. Depth synthesis and local warps for plausible image-based navigation. *ACM Transactions on Graphics (TOG)*, 2013. 2

[8] Shenchang Eric Chen and Lance Williams. View interpolation for image synthesis. *SIGGRAPH*, 1993. 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. *ICCV*, 2021. 3, 4

[10] 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 Transactions on Graphics (ToG)*, 2015. 2

[11] Edilson De Aguiar, Carsten Stoll, Christian Theobalt, Naveed Ahmed, Hans-Peter Seidel, and Sebastian Thrun. Performance capture from sparse multi-view video. *SIGGRAPH*, 2008. 2

[12] Paul E Debevec, Camillo J Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach. *SIGGRAPH*, 1996. 2

[13] Boyang Deng, JP Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Neural articulated shape approximation. *ECCV*, 2020. 4

[14] Matthew Flagg, Atsushi Nakazawa, Qiushuang Zhang, Sing Bing Kang, Young Kee Ryu, Irfan Essa, and James M Rehg. Human video textures. In *Proceedings of the 2009 symposium on Interactive 3D graphics and games*, pages 199–206, 2009. 1

[15] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. *ICCV*, 2021. 2

[16] Steven J Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F Cohen. The lumigraph. *SIGGRAPH*, 1996. 2

[17] 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 Transactions on Graphics (TOG)*, 2019. 1, 2

[18] Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Real-time deep dynamic characters. *ACM Transactions on Graphics (TOG)*, 2021. 2

[19] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11046–11056, 2021. 2

[20] Peter Hedman and Johannes Kopf. Instant 3D photography. *ACM Transactions on Graphics (TOG)*, 37(4):1–12, 2018. 2

[21] Peter Hedman, Tobias Ritschel, George Drettakis, and Gabriel Brostow. Scalable inside-out image-based rendering. *ACM Transactions on Graphics (TOG)*, 2016. 2

[22] Peter Hedman, Pratul P. Srinivasan, Ben Mildenhall, Jonathan T. Barron, and Paul Debevec. Baking neural radiance fields for real-time view synthesis. *ICCV*, 2021. 2

[23] Amir Hertz, Or Perel, Raja Giryes, Olga Sorkine-Hornung, and Daniel Cohen-Or. SAPE: Spatially-adaptive progressive encoding for neural optimization. *arXiv:2104.09125*, 2021. 4

[24] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. *CVPR*, 2020. 2, 4

[25] Zhang Jiakai, Liu Xinhang, Ye Xinyi, Zhao Fuqiang, Zhang Yanshun, Wu Minye, Zhang Yingliang, Xu Lan, and Yu Jingyi. Editable free-viewpoint video using a layered neural representation. *ACM SIGGRAPH*, 2021. 2

[26] Takeo Kanade, Peter Rander, and PJ Narayanan. Virtualized reality: Constructing virtual worlds from real scenes. *IEEE multimedia*, 1997. 1, 2

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

[28] Nikos Kolotouros, Georgios Pavlakos, Michael J. Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. *ICCV*, 2019. 5

[29] Marc Levoy and Pat Hanrahan. Light field rendering. *SIGGRAPH*, 1996. 2

[30] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. *CVPR*, 2021. 2

[31] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Trans. Graph.(ACM SIGGRAPH Asia)*, 2021. 2

[32] Lingjie Liu, Weipeng Xu, Marc Habermann, Michael Zollhöfer, Florian Bernard, Hyeongwoo Kim, Wenping Wang, and Christian Theobalt. Neural human video rendering by learning dynamic textures and rendering-to-video translation. *IEEE Transactions on Visualization and Computer Graphics*, 2020. 1, 2

[33] 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)*, 2015. 1, 2

[34] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. *arXiv:1705.09368*, 2017. 2

[35] 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. *ACM Transactions on Graphics (TOG)*, 2018. 1, 2

[36] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the wild: Neural radiance fields for unconstrained photo collections. *CVPR*, 2021. 1

[37] Wojciech Matusik, Chris Buehler, Ramesh Raskar, Steven J Gortler, and Leonard McMillan. Image-based visual hulls. *SIGGRAPH*, 2000. 1, 2

[38] Nelson Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1995. 4

[39] Marko Mihajlovic, Yan Zhang, Michael J Black, and Siyu Tang. LEAP: Learning articulated occupancy of people. *CVPR*, 2021. 4

[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. *ECCV*, 2020. 2, 3, 4, 5

[41] Natalia Neverova, Riza Alp Guler, and Iasonas Kokkinos. Dense pose transfer. *ECCV*, 2018. 2

[42] Michael Niemeyer and Andreas Geiger. GIRAFFE: Representing scenes as compositional generative neural feature fields. *CVPR*, 2021. 2

[43] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. *ICCV*, 2021. 2

[44] Rohit Pandey, Anastasia Tkach, Shuoran Yang, Pavel Pidlypenskyi, Jonathan Taylor, Ricardo Martin-Brualla, Andrea Tagliasacchi, George Papandreou, Philip Davidson, Cem Keskin, et al. Volumetric capture of humans with a single RGBD camera via semi-parametric learning. *CVPR*, 2019. 2

[45] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. *ICCV*, 2021. 1, 2, 4

[46] 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. *SIGGRAPH Asia*, 2021. 1, 2, 5, 7

[47] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. *ICCV*, 2021. 1, 2, 4

[48] 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. *CVPR*, 2021. 1, 2, 5, 7

[49] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural radiance fields for dynamic scenes. *CVPR*, 2020. 1, 2

[50] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J. Black. SCANimate: Weakly supervised learning of skinned clothed avatar networks. *CVPR*, 2021. 4

[51] Soubhik Sanyal, Alex Vorobiov, Timo Bolkart, Matthew Loper, Betty Mohler, Larry S Davis, Javier Romero, and Michael J Black. Learning realistic human reposing using cyclic self-supervision with 3D shape, pose, and appearance consistency. In *CVPR*, 2021. 2

[52] Kripasindhu Sarkar, Vladislav Golyanik, Lingjie Liu, and Christian Theobalt. Style and pose control for image synthesis of humans from a single monocular view. *arXiv preprint arXiv:2102.11263*, 2021. 2

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

[54] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. GRAF: Generative radiance fields for 3d-aware image synthesis. *NeurIPS*, 2020. 5

[55] Harry Shum and Sing Bing Kang. Review of image-based rendering techniques. *Visual Communications and Image Processing 2000*, 2000. 2

[56] 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. *CVPR*, 2021. 2

[57] Jonathan Starck, Gregor Miller, and Adrian Hilton. Video-based character animation. *ACM SIGGRAPH/Eurographics symposium on Computer animation*, 2005. 1, 2

[58] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. In *Advances in Neural Information Processing Systems*, 2021. 2

[59] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010. 2

[60] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *NeurIPS*, 2020. 2, 4

[61] Garvita Tiwari, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. Neural-gif: Neural generalized implicit functions for animating people in clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11708–11718, 2021. 4

[62] 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. *ICCV*, 2021. 1, 2

[63] Daniel Vlasic, Ilya Baran, Wojciech Matusik, and Jovan Popović. Articulated mesh animation from multi-view silhouettes. *SIGGRAPH*, 2008. 1, 2

[64] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. IBRNet: Learning multi-view image-based rendering. *CVPR*, 2021. 1

[65] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *NeurIPS*, 2018. 2

[66] Tuanfeng Y Wang, Duygu Ceylan, Krishna Kumar Singh, and Niloy J Mitra. Dance in the wild: Monocular human animation with neural dynamic appearance synthesis. In *2021 International Conference on 3D Vision (3DV)*, pages 268–277. IEEE, 2021. 2

[67] Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. Vid2Actor: Free-viewpoint animatable person synthesis from video in the wild. *arXiv:2012.12884*, 2020. 2, 3, 4

[68] Minye Wu, Yuehao Wang, Qiang Hu, and Jingyi Yu. Multi-view neural human rendering. *CVPR*, 2020. 2

[69] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. *CVPR*, 2021. 2

[70] Feng Xu, Yebin Liu, Carsten Stoll, James Tompkin, Gaurav Bharaj, Qionghai Dai, Hans-Peter Seidel, Jan Kautz, and Christian Theobalt. Video-based characters: creating new human performances from a multi-view video database. *SIGGRAPH*, 2011. 2

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

[72] Ze Yang, Shenlong Wang, Siva Manivasagam, Zeng Huang, Wei-Chiu Ma, Xinchen Yan, Ersin Yumer, and Raquel Urtasun. S3: Neural shape, skeleton, and skinning fields for 3d human modeling. *CVPR*, 2021. 2, 4

[73] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. NeRF++: Analyzing and improving neural radiance fields. *arXiv:2010.07492*, 2020. 1, 2

[74] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *CVPR*, 2018. 5, 7

[75] 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. *SIGGRAPH Asia*, 2021. 2

[76] C Lawrence Zitnick, Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High-quality video view interpolation using a layered representation. *ACM transactions on graphics (TOG)*, 2004. 2