MoRF: Mobile Realistic Fullbody Avatars from a Monocular Video

Renat Bashirov[1] Alexey Larionov[1] Evgeniya Ustinova[1] Mikhail Sidorenko[1]
David Svitov[1] Ilya Zakharkin[2*] Victor Lempitsky[3*]
[1]Samsung Research [2]ZERO10 [3]Cinemersive Labs

Abstract

*We present a system to create **Mo**bile **R**ealistic **F**ullbody (MoRF) avatars. MoRF avatars are rendered in real-time on mobile devices, learned from monocular videos, and have high realism. We use SMPL-**X** as a proxy geometry and render it with DNR (neural texture and image-2-image network). We improve on prior work, by overfitting per-frame warping fields in the neural texture space, allowing to better align the training signal between different frames. We also refine SMPL-X mesh fitting procedure to improve the overall avatar quality. In the comparisons to other monocular video-based avatar systems, MoRF avatars achieve higher image sharpness and temporal consistency. Participants of our user study also preferred avatars generated by MoRF.*

1. Introduction

This work focuses on fullbody human avatars with fast rendering, realism and the ease of acquisition. For our system we: 1) use short (*i.e.* minute-long) monocular videos as input, 2) train for a few hours on a single GPU, 3) provide sharp texture of the resulting avatar, and 4) can render in real-time on mobile devices. As far as we know, despite impressive progress, existing systems fall short of these criteria. Specifically, some rely on multi-view videos [39, 52], or a depth sensor [34, 62], some leverage implicit representations whose rendering process is far from real-time especially on mobile devices [5, 21, 36, 42, 48, 49], whilst others take days to converge on multiple GPUs [64, 68].

We therefore present a system for **Mo**bile **R**ealistic **F**ullbody (MoRF) avatars. Our avatars are realistic and exhibit temporal consistency of generated images, and they can be rendered in realtime at 30 FPS on mobile devices (Fig. 1) with Qualcomm Snapdragon 888 SoC (2020's flagship SoC for mobile phones). Furthermore, they can be acquired from monocular videos in a few hours on a single NVIDIA RTX 3090 GPU. We rely on the deferred neural rendering (DNR) idea [60], like in numerous prior avatar systems [53, 67–69], to achieve rendering speed and realism. DNR represents 3D objects using coarse 3D geometry and *neural textures*, which are supplemented by a rendering image-2-image translation network to produce realistic images. For geometric modeling, we use the well-established SMPL-X body model [47].

Figure 1. Three MoRF avatars (and a real person in gray) rendered on a mobile phone in realtime (at least 30 frames per second) in the augmented reality mode. Each of the avatars was created from a monocular video, while the poses and the camera positions of these examples differ from the training frames.

We discovered that naively implementing the DNR approach results in underfitting blurriness or overfitting flickering. To address them, we introduce a new component into the DNR learning process that we call the *neural texture warping*. For each training image/frame, it fits a warping field for the neural texture, thus compensating between-frame mesh fit misalignment and facilitating the task for image-2-image UNet renderer.

The warping fields are only used for training to compensate for mesh fitting imperfections, this produces sharp and static texture for the resulting avatar. *We don't solve the task*

*Contributed during employment at Samsung Research

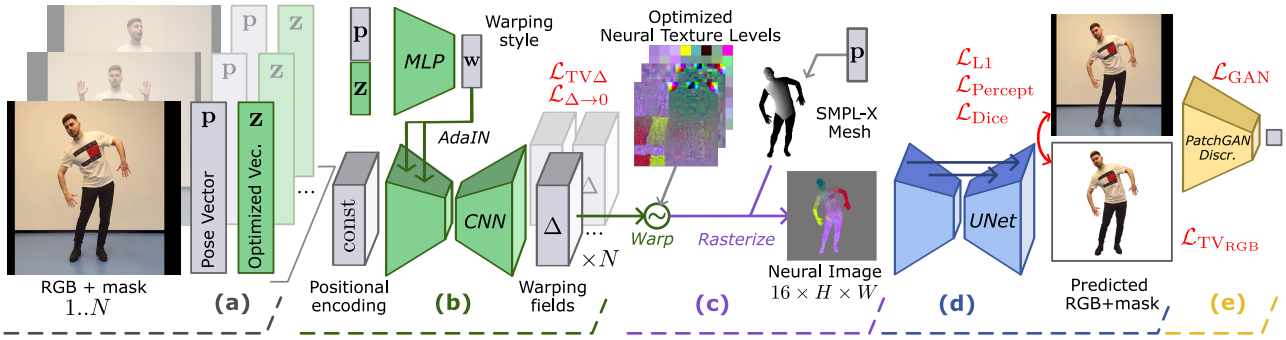


Figure 2. The proposed architecture. The colored blocks highlight end-to-end optimized parts. (a) The frames, pose vectors, and segmentation masks (not shown) are all inputs to the modeling process. In addition, for each frame, a latent texture warping vector is introduced. (b) A texture warping field is computed for each frame using a convolutional architecture and the respective warping vector. (c) The neural texture associated with the avatar is warped according to the predicted field and is superimposed over a posed SMPL-X mesh. (d) The rendering network predicts an RGB image as well as a segmentation mask. This network also inpaints out-of-mesh details, *e.g.* garments and hair. (e) We supervise with pixelwise, perceptual, adversarial losses on predicted images; Dice loss on predicted segmentations.

of clothing modelling/movement and the warping fields are discarded after the training process has converged.

We evaluate our approach on 3 datasets: self-captured, ZJU-MoCap [13, 49] and People Snapshot [3]. Self-captured and ZJU-MoCap provide ground truth for novel body and camera poses respectively. We also conduct user studies for avatar rendering quality and for mesh fitting refinements ablation. We compare against InstantAvatar [26], HF-Avatar [71], Anim-NeRF [48], HumanNeRF [66], StylePeople [17], ANR [53] and NeuMan [27].

2. Related Work

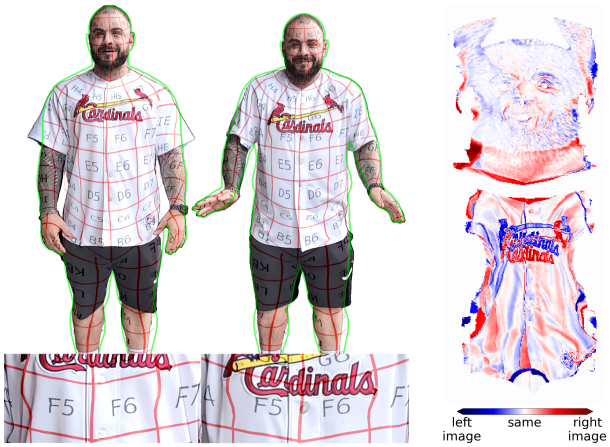
Modeling geometry of dynamic non-rigid scenes is considered in a number of recent approaches either for capturing human actors [6, 18, 20, 23, 33, 35, 36, 40, 61, 66] or general scenes [38, 46, 51]. Towards this end, 3D scans are required for supervision in [20, 23, 35, 40, 61] to learn rigged human geometry. Likewise, [6, 18, 33, 36, 38, 50, 54] utilize multi-view data to capture appearance and produce photo-realistic avatars under arbitrary viewpoints and in arbitrary poses. Several methods [46, 51, 66] use monocular videos but allow free-viewpoint rendering only.

Some of the monocular fullbody avatar methods model the human geometry implicitly [36, 48, 66], others output the classical mesh+texture format [1, 2, 4, 17, 53, 59, 71]. Similarly to StylePeople [17] and ANR [53], our method is based on DNR [60] allowing for efficient convolutional rendering on a mobile device. However, StylePeople leaves modeling of geometry imperfections and frame-to-frame misalignments to the neural renderer, while ANR handles these issues by limiting the number of frames used for the neural texture optimization. HF-Avatar [71] uses the synthetic images as supervision for the neural texture, because of their consistency and good alignment with the geometry. [2] proposes a texture merging optimization procedure to prevent texture averaging among different views. Our neural texture warping allows for consistent learning of the neural texture without training on synthetic data [71], excluding most frames from texture learning [53], or overfitting the neural renderer to unmodeled variations [17].

Many of the mentioned works introduce learned submodules [18, 33, 35, 36, 38, 61, 66, 71] or pipeline stages [2, 4] to estimate additional geometry warping in 3D canonical (unposed or view-agnostic) space. Likewise, [33, 35, 36, 61, 66] parametrize the 3D warping fields or displacements [40, 71] as neural networks conditioned on human pose to facilitate learning of complex pose-dependent geometry. Methods for general scenes leverage other variants of conditioning. For example [46, 51], that rely on monocular videos, condition 3D warping fields on the point in time and learnable latent of a training frame respectively, while [38] introduce view-conditional warpings as they use multi-view capture. The way we introduce warping in this work is different to all above-referenced approaches, as our learnable warping operates in the canonical 2D texture space. Operating in the texture space naturally fits the DNR approach that we are based upon.

3. Method

Our approach aims to learn a rigged full-body human avatar using a relatively short (e.g. a minute-long) monocular video. We begin by briefly describing deferred neural rendering framework, as it is at the core of our system. We then introduce the neural texture warping that facilitates the training process. We also mention important architectural decisions that contribute to the avatars quality.



(a) Image space (b) Texture space

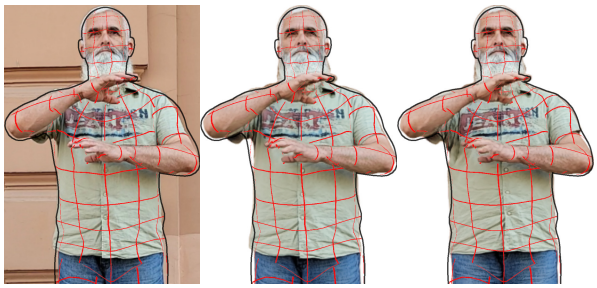
Figure 3. (a) Mesh surfaces (represented by red lines) fitted to two different frames are located inconsistently. Because of the rendering network's overfitting tendency, such misalignment can cause artifacts of the learned avatar. (b) A partial plot of $\Pi_{\mathbf{T}}(\mathbf{I}_2) - \Pi_{\mathbf{T}}(\mathbf{I}_1)$, where $\Pi_{\mathbf{T}}$ is the texture space projection. The red and blue colors correspond to the two poses from (a), whilst the white color indicates that the colors match. The texture space is clearly misaligned.

3.1. Recap: Deferred Neural Rendering

Deferred neural rendering (DNR) [60] models scenes by combining a geometric proxy defined by a triangular mesh, a neural texture $\mathbf{T}$ (with C channels), and an image-to-image rendering network (renderer) $\mathcal{R}$ with a convolutional architecture. Novel views for certain camera parameters are synthesized in two steps. The first is to project the mesh onto the camera view, while superimposing the texture on the mesh using texture mapping. The resulting C-channeled image can then be "translated" into an RGB image (plus optionally a mask) by the rendering network. During training/fitting to a given dataset of images with known corresponding camera parameters, the latent texture $\mathbf{T}$ and the parameters of the rendering network $\mathcal{R}$ are jointly optimized to minimize losses between training images and images predicted by $\mathcal{R}$.

3.2. Neural Texture Warping

The notable finding of [60] is that the quality of deferred neural rendering degrades very gracefully in the presence of coarseness or systematic errors in the geometric proxy. This has allowed several recent works [53, 67–69] to apply DNR to fullbody modeling using SMPL-X as an underlying geometry, despite the big gap between this geometric proxy and the actual avatar content that might include sliding or loose clothing. DNR based on SMPL-X geometry essentially "off-loads" clothing and hair modeling to the rendering network that operates in the 2D domain. While appealing for its simplicity, we discovered that such an approach frequently fails to fit the training data, because frames with similar camera parameters and body pose can differ in hair geometry or clothing draping (see Figure 3), which the DNR approach cannot model well. We propose modeling them in the texture space with 2D warping maps (see Figure 2). Our cornerstone is that the texture space by construction is a canonical space for the human body – it is invariant to body poses and camera views. At the same time it covers the whole 3D body surface. As a result, our warping approach is less ambiguous than commonly applied warping in the 3D spatial domain, and arguably is easier to train from monocular videos.



(a) (b) (c)

Figure 4. (a) The SMPL-X (shown by red lines) superimposed on a training frame. (b) The corresponding **avatar** image produced by MoRF, that aligns with the training image well as compared to (c) where the learned texture warping field of the given frame was replaced with a zero warping field. Note, that the clothes appear in a *neutral* position: the centralized shirt, the undeformed hemline. As seen, the warping field successfully "absorbs" the unique cloth deformation on this training frame, allowing the neural texture and the rendering network to learn an easier task.

For each training frame, we learn a frame-specific warping field for the neural texture. The frame specific information is represented by a combination of the frame's SMPL-X pose vector $\mathbf{p}$ and a latent $\mathbf{z}$. Those are passed through a multi-layered perceptron (MLP) $\mathcal{W}$ to obtain a style vector $\mathbf{w}$. We predict the warping fields with an encoder-decoder convolutional architecture $\mathcal{E} \odot \mathcal{F}$, the input to which is a constant tensor of positional encodings of the texture coordinates. The style vector $\mathbf{w}$ conditions the encoder part via the Adaptive Instance Normalization blocks [22]. A predicted warping field is a 2-channel tensor, that specifies per-texel offsets of texture coordinates. See supp. mat. for the exact details of the architectures $\mathcal{E}$, $\mathcal{F}$, $\mathcal{W}$. Figure 4 shows the effect of the learned warping fields.

During model fitting, we use gradient descent to fit per-frame latent vectors $\mathbf{z}$, and parameters of the warping networks $\mathcal{E}$, $\mathcal{F}$, $\mathcal{W}$, the neural texture $\mathbf{T}$ and the rendering network $\mathcal{R}$, which are shared for all training frames. The latent vectors $\mathbf{z}$ encode the variations in avatar geometry that were

(a) (b) (c) (d)

Figure 5. Results of MoRF on novel poses, for two neural texture initialization methods: (a)/(c) - random initialization, (b)/(d) - spectral initialization. Some body parts may exhibit overfitting artifacts. Local smoothness of spectral initialization acts as a prior to solve that.

not captured by the underlying SMPL-X model, and the frame-specific imperfections of SMPL-X fitting (Fig. 4).

For rendering in novel poses, we compute the warping field corresponding to the neutral A-pose $\mathbf{p}$ and the average latent vector $\mathbf{z}$. Such warping field is applied to the neural texture once. First, such an approach is suitable in the case of limited train set. Second, this helps eliminating the need for warping computations during inference, thus preserving high rendering speed.

Texture initialization We found that, as compared to random initialization, better results can be achieved with *spectral initialization* of the neural texture as in StylePeople [17]: each texel value is initialized with spectral coordinates of the corresponding point on the SMPL-X mesh graph (see examples in supp. mat.) We learn neural textures with small learning rates to avoid vanishing of the spectral information. We found that the use of spectral initialization was more pronounced in our system compared to [17]. In particular, this initialization has minor effect on image fidelity, but it significantly slows down overfitting of the rendering network, giving the warping network time to converge and facilitating temporal consistency. It also contributes in plausible inpainting of body parts not covered in the input video (see Figure 5).

4. Implementation details

4.1. Mesh fitting

For both our system and the standard DNR model to perform well, precise and time-consistent SMPL-X mesh fitting to the training video frames (and at test time) is important. As for RGB photos, the convolutional regression-based techniques [10, 14, 28, 30, 31, 57, 58] produce somewhat accurate but inconsistent results that frequently mismatch the silhouettes. On the contrary, results of iterative non-linear optimization [7] better correspond 2D RGB images but are less plausible or unstable in 3D, since the fitting process relies on 2D image keypoints, which are poorly constrained along the depth dimension. Regression-based and optimization-based methods are unable to handle complex poses. Below we describe a few enhancements to SMPL-X optimization-based fitting that increase the stability of DNR-based avatar systems such as ours. We later validate the individual impact of these adjustments through ablation study.

Shared shape A person's identity does not change between frames of a training video, thus the shape-specific parameters of SMPL-X model can be shared.

Silhouette loss We perform differentiable rendering of SMPL-X meshes via `nvdiffrast` [32]. We then apply Dice [43] loss to pull rendered SMPL-X silhouettes towards estimations of a semantic segmentation network. In the presence of loose clothing, the inclusion of this term tends to "fatten" the human shape.

Temporal loss The non-linear optimization-based fitting method tends to flicker on side views when the person's farther side is not visible. To compensate for this effect, we minimize the L2 distance between adjacent body pose vectors (in axis-angle format). This term is efficient for our scenario of training videos with slow body movements. Applying it to "in-the-wild" videos with fast motions may be ineffective or even harmful.

Automatic frame filtering The mesh fitting procedure is vulnerable to quick movements and depth ambiguity even with temporal guidance. Some training frames may be estimated with significantly incorrect poses. This causes a conflict between the 3D mesh surface and the neural texture, resulting in the appearance of "ghost" limbs and flickering. We use a simple heuristic that we describe in supp. mat. to automatically filter out bad frames (typically, less than 5% of frames are discarded).

4.2. Architectural details

Training As a rendering network, we use U-Net [55] (a total of ≈14.4M parameters). Its decoder is based on bilinear upsampling blocks. We provide the definitions of neural layers in supp. mat. We use ADAM [29] to optimize the architecture from scratch for a target person. On a single RTX 3090 GPU, convergence to plateau takes 8 hours. Our preliminary results suggest that faster acquisition is possible, if the architecture is pretrained on a dataset of people, and then fine-tuned to a target person. Below, however, we report on experiments that fit the model to a target video from scratch.

Ad-hoc inpainting of unseen body parts Our minute-long training videos rarely include all body parts of the target person. Often top of the head, soles, and armpits have

no ground truth. Those parts can be blurred on avatar renders, or worse – excluded from the predicted segmentation. We overcome this by post-train finetuning for areas of the neural texture that correspond to unseen body parts. We first generate a pseudo ground-truth by projecting training frames into an RGB texture using SMPL-X fits, then we fill gaps on the RGB texture with nearby color, and render new ground-truth with extreme body and camera poses that show unseen body parts. We then fine-tune with L1 per-pixel loss.

Inputs We use OpenPose [8] to detect 2D keypoints, which is required for mesh fitting. We retrain Graphonomy [15] segmentation network on CCIHP dataset [37], which includes accessories like watches, belts, and glasses, etc. Meanwhile the publicly available Graphonomy trained on CIHP dataset predicts those as background.

4.3. Loss functions

For MoRF training, we use popular loss functions from previous works [17, 60, 63] with modifications. We further elaborate on our learning objective:

$$\mathcal{L}_{\text{total}} = \sum_{i \in \mathfrak{L}} \lambda_i \mathcal{L}_i \qquad (1)$$
$$\mathfrak{L} = \{\text{L1}, \text{Percept}, \text{GAN}, \text{Dice}, \text{TV}_{\text{RGB}}, \text{TV}_{\Delta}, \Delta \to 0\}$$

Image losses These include L1 per-pixel loss $\mathcal{L}_{\text{L1}}$ and the VGG [56] based perceptual loss [12] $\mathcal{L}_{\text{Percept}}$. To enforce the realism of the rendered images, we utilized the non-saturating adversarial loss [16] $\mathcal{L}_{\text{GAN}}$ on predictions of multi-scale patch-based discriminators (PatchGAN from Isola *et al*. [25]).

Segmentation loss In our case, the estimated body mesh (SMPL-X) is coarse and frequently misaligned with the underlying human content. To assist the renderer in generating content outside of the rasterized mesh, the renderer predicts the human mask in addition to the RGB image. We supervise the masks on the pseudo ground truth masks (see paragraph 4.2), using Dice [43] $\mathcal{L}_{\text{Dice}}$ segmentation loss.

Regularization For the warping fields, we penalize L2 norm ($\mathcal{L}_{\Delta \to 0}$) to prevent large deviations from identity warping, and also apply total-variation (TV) loss [9] $\mathcal{L}_{\text{TV}_\Delta}$ for local smoothness. We use a low weighted TV loss on generated images $\mathcal{L}_{\text{TV}_{\text{RGB}}}$ to slightly improve smoothness.

4.4. Avatar rendering on a mobile device

We developed a telepresence mobile app for Qualcomm Snapdragon SoC using Android, ARCore [45], and OpenGL. Our method can run natively at resolution of 640×640 pixels and above 30 frames per second (measured on Qualcomm Snapdragon 888, see Fig. 1). We use OpenGL shaders to generate input neural rasterizations and to infer posed SMPL-X meshes from body joint rotations. We stumbled upon OpenGL limit on the depth of the neural texture channels. We were able to pack at most 16 channels of neural rasterizations into OpenGL framebuffers, using quantization [44] into eight-bit integers. Similarly, we use post-training quantization on the weights of the rendering network. This enables the network to run at high speeds on the Qualcomm Digital Signal Processor. Quality degradation due to quantization is relatively low in our system. Because single-person avatars have low variance in appearance, the numerical range of neural network activations is small, which is advantageous for quantization.

5. Experiments

We compare our method to similar recent approaches and provide an ablation study for our mesh fitting adjustments. We conduct quantitative comparisons as well as user study comparisons. We provide sample side-by-side comparisons of the methods in supp. mat.

5.1. Self-captured dataset

We have gathered a dataset of stationary monocular videos of ten people with different demographics, body shapes, and clothes topology. We note that despite improvements in mesh fitting, it is still imperfect, therefore we captured the training videos with a specific scenario that avoids complex poses. In one-minute training videos people move slowly, turn 360°, and show their hands in front of their torso. The one-minute hold-out videos include mostly novel poses not present in training videos, namely various arm and hip movements. See samples of training and hold-out data in supp. mat.

Table 1. Quantitative metrics and the results of the user preference study, in which each method is compared against MoRF. See the text for the discussion of the results.

	SSIM $\cdot 10^2 \uparrow$	LPIPS $\cdot 10^2 \downarrow$	tLPIPS $\cdot 10^3 \downarrow$	% of users preferred **MoRF**
Scenario A: 512×512 *px, training only on A-poses*				
InstantAvatar	92.52	7.91	7.52	96.5% ↑
HF-Avatar	93.0	7.25	4.00	90.0% ↑
AnimNeRF	92.7	6.74	2.96	98.0% ↑
MoRF	94.5	3.95	2.56	N/A
Scenario B: 640×640 *px, full training sequences*				
HumanNeRF	87.4	13.00	19.3	95.0% ↑
StylePeople	90.0	6.13	4.77	60.5% ↑
ANR	90.0	6.14	4.76	69.2% ↑
MoRF	89.9	6.23	4.66	N/A

We use publicly available implementations of InstantAvatar, HF-Avatar, Anim-NeRF, NeuMan and HumanNeRF. To compare with InstantAvatar, HF-Avatar, Anim-NeRF methods, we exclude the sections of training

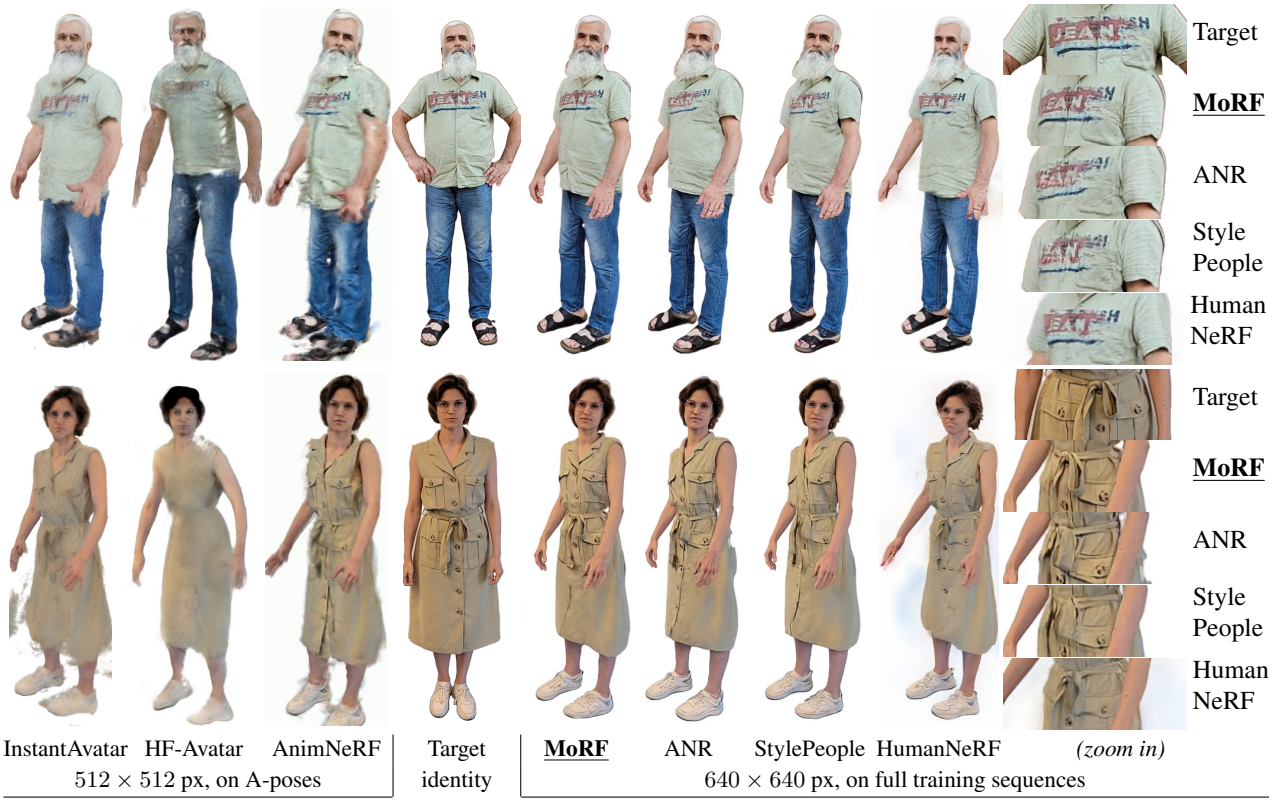


Figure 6. Avatars created with different monocular video-based approaches. MoRF achieves the highest visual quality in this comparison. See text for more details.

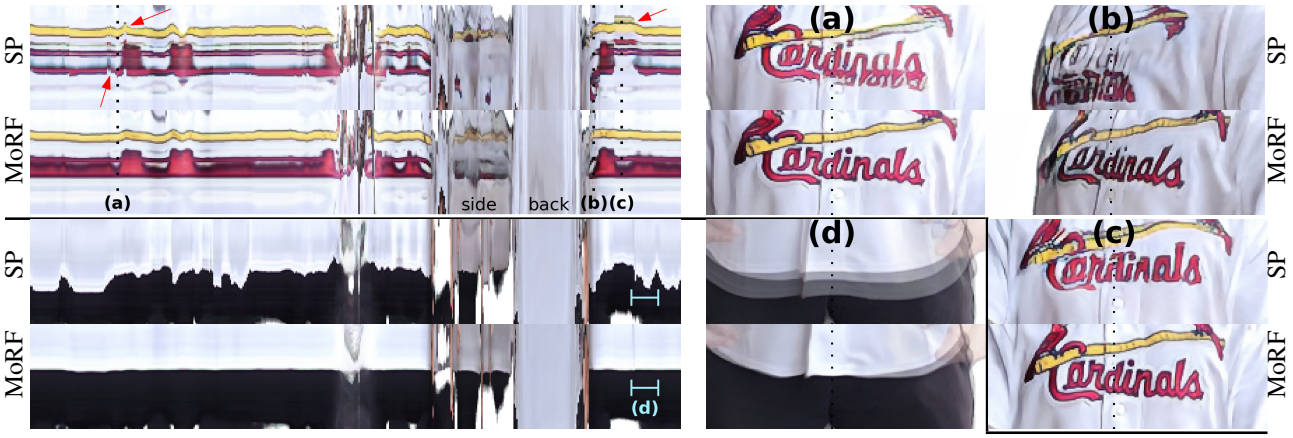


Figure 7. To illustrate the difference in temporal consistency, we show the space-time y-t slices (left). In the case of StylePeople (SP), the T-shirt pattern (top plots) abruptly tears vertically (accentuated with red arrows), resulting in (a) flickering, (b) "ghost" patterns, (c) poor clarity, and loss of pattern rendering for rotated human body (see the section labeled *"side"*). These artifacts are absent in our approach's results, which also achieves good temporal stability on the hemline (bottom plots). (d) Shows an average image over a 20-frame interval (colored in light-blue).

videos where people show their hands in front of their torso (leaving about half of the frames), this forms *Scenario A* comparison. Anim-NeRF limits the resolution for *Scenario A* to 512×512 px. HF-Avatar's mesh fitting adequately performed for training sequences, but struggled for hold-out sequences, therefore for them we used our

Table 2. Mesh fitting used in training and validation.

	Training	Hold-out
InstantAvatar	their SMPL	their SMPL
HF-Avatar	their SMPL	our SMPL & their train body shape
AnimNeRF	their SMPL	their SMPL
HumanNeRF	our SMPL	our SMPL
StylePeople	our SMPL-X	our SMPL-X
ANR	our SMPL-X	our SMPL-X
MoRF	our SMPL-X	our SMPL-X

modified mesh fitting. HumanNeRF implementation does not include its own mesh fitting method, thus we use ours instead. See Table 2 for an overview of mesh fitting methods applied in our experiments.

We compare against HumanNeRF, ANR and StylePeople trained on full uncut videos at 640×640 px resolution – the maximum resolution at which we can run 30 FPS on a mobile device. This forms *Scenario B* comparison. For StylePeople, we use all the proposed components as per [17, 60], but take our segmentation, mesh fitting, losses, spectral texture initialization and architectures of the rendering and discriminator networks. In the comparison table, StylePeople baseline differs from MoRF only by the absence of the neural texture warping, in order to compare only novelty part of our work. To compare with ANR, we re-implement their split optimization scheme and add it to the StylePeople baseline (see supp. mat. for more details).

To quantify visual quality, we use three metrics, computed between real and generated images. We employ two image-based metrics: SSIM [65] and LPIPS [70]. To measure temporal stability, we use tLPIPS [11].

We conduct a *user study* (the right-most column of Table 1) in which we show pairs of sequences, one of which is the MoRF result (in randomized order). We took novel body motion sequences from AMASS [41] dataset. We asked people to indicate their preference, and thus obtain the percentage of responses when MoRF was preferred over each of the other methods.

The results are shown in Table 1. In our comparison, AnimNeRF and InstantAvatar produce NeRF blob artifacts. HumanNeRF similarly exhibits NeRF artifacts in motion, and its training diverged for some training videos, with the reasons unrelated to MoRF's mesh fitting quality (see supp. mat. for evaluation scores and images). Due to such divergence, HumanNeRF has very high values for LPIPS and tLPIPS metrics. Non-NeRF HF-Avatar method suffers from significant blurring. In comparison to StylePeople and ANR, our approach outperforms both of them in temporal stability while slightly underperforming in SSIM and LPIPS metrics. We explain it by noticing, that when turning off warping fields on inference, the rendered avatar's texture appears in a neutral position (Fig. 4), which is almost certainly misaligned with the real dynamic garments from the hold-out ground truth. Thus, the metrics capture mostly pixel-wise discrepancy between garments, which might not match the human perception of pleasant avatars. Qualitatively, we observe at least as good image quality for MoRF as that of StylePeople and ANR. The advantage of our method is validated by the user study on the videos, where MoRF is preferred over other methods. We also provide a qualitative comparison of MoRF with StylePeople in our mobile AR app in supp. mat.

In Figure 6 we show the qualitative comparisons between the methods. In Figure 7, we further compare the temporal stability of MoRF avatars with StylePeople avatars using time-space slices. Here we created avatar images with a novel animation sequence and centered them on a single SMPL-X mesh vertex: on a T-shirt pattern, or the hemline. We demonstrate extreme failure cases for StylePeople that do not occur with MoRF. As seen, our method provides good vertical stability of rendered parts, confirming that the tasks of mesh misalignment absorption and translation for latent neural renderings into color images are disentangled.

Mesh fitting ablation study In Table 3, we perform an ablation study on mesh fitting modifications described in Section 4.1. We start with the SMPLify-X baseline [7] and add the improvements one-by-one. We retrain the MoRF avatars for the new fits, and perform a user study on pairs of videos: one is the baseline, the other is a modification. Every pair is shown to 25 people, one person judges at most 6 pairs every 6 hours. Although on average any modification is favored by the user study over SMPLify-X, on particular avatar subjects the score can significantly drop, *e.g.* as low as 12% for the temporal loss. However, when all modifications are used, the lowest preference score among all subjects was 52%, indicating the consistently positive impact of the proposed fitting improvements on the avatar quality.

Table 3. Ablation user study on the mesh fitting procedure. We report a percentage of users preferring avatars trained on modified mesh fits over baseline SMPLify-X [7, 47].

SMPLify-X	+Shared shape	+Frames filter	+Temporal loss	+Silhouette loss	+All
N/A	57.6%	58.0%	64.4%	84.4%	84.8%

5.2. ZJU-MoCap

To evalute on ZJU-MoCap dataset, we follow the protocol/split from HumanNeRF and compare to HumanNeRF and StylePeople with two types of SMPL mesh fits: type A) the original multi-view fits from ZJU-MoCap, and type B) our monocular fitting results. For both types of fits, HumanNeRF exhibits ghosting and NeRF blob artifacts. Despite MoCap origin of this dataset, it's original SMPL multi-

	Subject 377			Subject 394		
	PSNR↑	SSIM↑	LPIPS↓∗	PSNR↑	SSIM↑	LPIPS↓∗
	ZJU-MoCap multi-view SMPL fits (type A)					
HumanNeRF	28.54	0.9557	31.203	28.67	0.9448	43.980
StylePeople	29.30	0.9639	29.865	28.93	0.9472	40.828
MoRF	29.03	0.9629	31.606	29.13	0.9483	41.014
	Our single-view SMPL fits (type B)					
HumanNeRF	27.74	0.9572	38.448	27.99	0.9426	51.120
StylePeople	27.25	0.9571	42.021	27.92	0.9440	47.956
MoRF	27.32	0.9573	42.156	27.88	0.9442	47.692

Table 4. Evaluation on ZJU-MoCap dataset. (∗) LPIPS $\times 10^3$.

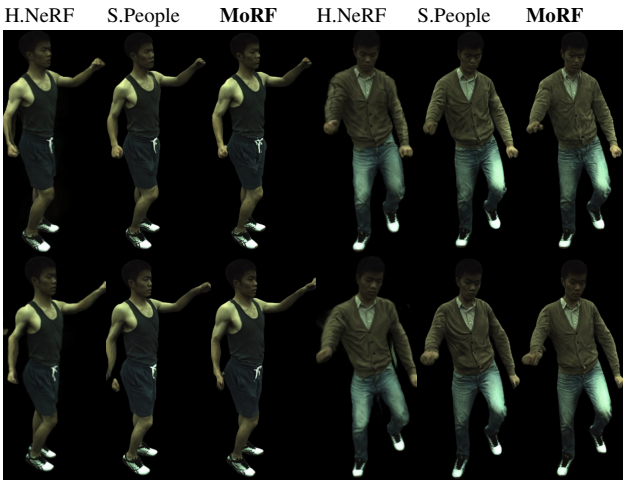


Figure 8. Results on ZJU-MoCap dataset (renderings of novel views far from the train view). Avatars of subjects 377 *(left)*, 394 *(right)* are shown, trained on multi-view *(top)* or on our monocular *(bottom)* mesh fits. Zoom-in is highly recommended.

view fits (type A) are less aligned with ground truth images than our monocular fits (type B). As a result, given type A fits, we find out that MoRF performs worse that StylePeople baseline (Tab. 4, Fig. 8 – zoom-in highly recommended) because neutral texture warping could not converge. On the contrary, given well aligned type B mesh fits, MoRF produces sharper output renderings and less flicker. This sets the application scope of our rendering method: the expected effect is achieved with our mesh fitting and capture scenario.

5.3. People Snapshot

We have also found that MoRF works well on **People Snapshot** data. See the comparison on two idetities from People Snapshot with InstantAvatar and StylePeople in Fig. 9.

5.4. Experiments analyses

For our self-captured dataset, all NeRF-based methods (InstantAvatar, Anim-NeRF and HumanNeRF) produce noticeable NeRF blob artifacts, and HumanNeRF has the least of them.

Figure 9. Results on PeopleSnapshot produced by InstantAvatar *(left)*, StylePeople *(middle)* and MoRF *(right)*. The least texture averaging is demonstrated by MoRF.

Compared to NeRF-based methods, mesh-based rendering methods (StylePeople, ANR and MoRF) are more stable and do not produce blob artifacts. Moreover the inference speed for the fastest NeRF (InstantAvatar) is 15 FPS for 540×540 px on desktop GPU, compared to 30 FPS for 640×640 px on **mobile** GPU for MoRF.

Compared to StylePeople and ANR, MoRF can produce sharper and more temporally stable results given our well aligned mesh fits. This effect is noticeable on texture rich regions like on a T-shift print.

For the results of HF-Avatar see Figure 6. The NeuMan method struggles for static monocular videos like in our self-captured dataset, see supp. mat. for the exact details.

6. Discussion

We have presented a fullbody avatar system that has high realism, can be rendered in real-time on a mobile devices, and is learned from monocular videos. We evaluate our method on a self-captured dataset on novel poses, prove MoRF is faster and produces better rendering results compared to NeRFs. On ZJU-MoCap dataset, we find out that MoRF is sensitive to mesh fits quality. MoRF produces sharper and more temporally stable results (compared to StylePeople and ANR) given our well aligned mesh fits, the effect is noticeable on texture rich regions like on a T-shift.

The limitations of our method are: 1) MoRF does not model clothing movement or loose clothing, 2) there is no prior to render train-blind areas, 3) texture seams are not handled and warping artifacts could appear there, using spherical texture without seams is a future work.

References

[1] Thiemo Alldieck, Marcus Magnor, Bharat Lal Bhatnagar, Christian Theobalt, and Gerard Pons-Moll. Learning to reconstruct people in clothing from a single rgb camera. In *CVPR*, pages 1175–1186, 2019. 2

[2] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Detailed human avatars from monocular video. In *2018 International Conference on 3D Vision (3DV)*, pages 98–109. IEEE, 2018. 2

[3] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8387–8397, 2018. 2

[4] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. Video based reconstruction of 3d people models. In *CVPR*, pages 8387–8397, 2018. 2

[5] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *CVPR*, pages 1506–1515, 2022. 1

[6] Timur Bagautdinov, Chenglei Wu, Tomas Simon, Fabian Prada, Takaaki Shiratori, Shih-En Wei, Weipeng Xu, Yaser Sheikh, and Jason Saragih. Driving-signal aware full-body avatars. *ACM TOG*, 40(4):1–17, 2021. 2

[7] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *ECCV*, Lecture Notes in Computer Science. Springer International Publishing, Oct. 2016. 4, 7

[8] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE TPAMI*, 2019. 5

[9] Antonin Chambolle. An algorithm for total variation minimization and applications. *Journal of Mathematical imaging and vision*, 20:89–97, 2004. 5

[10] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Monocular expressive body regression through body-driven attention. In *ECCV*, pages 20–40. Springer, 2020. 4

[11] Mengyu Chu, You Xie, Laura Leal-Taixé, and Nils Thürey. Temporally coherent gans for video super-resolution (tecogan). *ArXiv*, abs/1811.09393, 2018. 7

[12] Alexey Dosovitskiy and Thomas Brox. Generating images with perceptual similarity metrics based on deep networks. *Advances in neural information processing systems*, 29, 2016. 5

[13] Qi Fang, Qing Shuai, Junting Dong, Hujun Bao, and Xiaowei Zhou. Reconstructing 3d human pose by watching humans in the mirror. In *CVPR*, 2021. 2

[14] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Collaborative regression of expressive bodies using moderation. In *2021 International Conference on 3D Vision (3DV)*, pages 792–804. IEEE, 2021. 4

[15] Ke Gong, Yiming Gao, Xiaodan Liang, Xiaohui Shen, Meng Wang, and Liang Lin. Graphonomy: Universal human parsing via graph transfer learning. In *CVPR*, pages 7450–7459, 2019. 5

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 5

[17] Artur Grigorev, Karim Iskakov, Anastasia Ianina, Renat Bashirov, Ilya Zakharkin, Alexander Vakhitov, and Victor Lempitsky. Stylepeople: A generative model of fullbody human avatars. In *CVPR*, pages 5151–5160, 2021. 2, 4, 5, 7, 14

[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)*, 40(4):1–16, 2021. 2

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 14

[20] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *ICCV*, pages 11046–11056, 2021. 2

[21] Fangzhou Hong, Zhaoxi Chen, Yushi Lan, Liang Pan, and Ziwei Liu. Eva3d: Compositional 3d human generation from 2d image collections. *arXiv preprint arXiv:2210.04888*, 2022. 1

[22] Xun Huang and Serge J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *CVPR*, pages 1510–1519, Oct. 2017. 3, 13

[23] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2020. 2

[24] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015. 13

[25] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 5, 13, 20

[26] 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*, pages 16922–16932, 2023. 2

[27] 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 (ECCV)*, 2022. 2, 12

[28] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, pages 7122–7131, 2018. 4

[29] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015. 4, 13

[30] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *CVPR*, pages 5253–5263, 2020. 4

[31] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, pages 2252–2261, 2019. 4

[32] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM TOG*, 39(6), 2020. 4

[33] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhöfer, Jürgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 419–436. Springer, 2022. 2

[34] Zhe Li, Tao Yu, Chuanyu Pan, Zerong Zheng, and Yebin Liu. Robust 3d self-portraits in seconds. In *CVPR*, pages 1344–1353, 2020. 1

[35] Zhe Li, Zerong Zheng, Hongwen Zhang, Chaonan Ji, and Yebin Liu. Avatarcap: Animatable avatar conditioned monocular human volumetric capture. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I*, pages 322–341. Springer, 2022. 2

[36] 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 TOG*, 40(6):1–16, 2021. 1, 2

[37] Angelique Loesch and Romaric Audigier. Describe me if you can! characterized instance-level human parsing. In *ICIP*, pages 2528–2532, 2021. 5

[38] 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. 2

[39] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. *ACM TOG*, 40(4):1–13, 2021. 1

[40] Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J Black. The power of points for modeling humans in clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10974–10984, 2021. 2

[41] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *CVPR*, pages 5442–5451, Oct. 2019. 7

[42] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1

[43] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571, 2016. 4, 5

[44] Markus Nagel, Marios Fournarakis, Rana Ali Amjad, Yelysei Bondarenko, Mart van Baalen, and Tijmen Blankevoort. A white paper on neural network quantization, 2021. 5

[45] Zainab Oufqir, Abdellatif El Abderrahmani, and Khalid Satori. Arkit and arcore in serve to augmented reality. In *2020 International Conference on Intelligent Systems and Computer Vision (ISCV)*, pages 1–7. IEEE, 2020. 5

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

[47] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, 2019. 1, 7, 13, 14

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

[49] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021. 1, 2

[50] Sergey Prokudin, Michael J Black, and Javier Romero. Smplpix: Neural avatars from 3d human models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1810–1819, 2021. 2

[51] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. *arXiv preprint arXiv:2011.13961*, 2020. 2

[52] Amit Raj, Umar Iqbal, Koki Nagano, Sameh Khamis, Pavlo Molchanov, James Hays, and Jan Kautz. Dracon–differentiable rasterization conditioned neural radiance fields for articulated avatars. *arXiv preprint arXiv:2203.15798*, 2022. 1

[53] Amit Raj, Julian Tanke, James Hays, Minh Vo, Carsten Stoll, and Christoph Lassner. Anr: Articulated neural rendering for virtual avatars. In *CVPR*, pages 3722–3731, 2021. 1, 2, 3, 12, 14

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

[55] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4, 14, 21

[56] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[57] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Black Michael J., and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, 2021. 4

[58] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3d people in depth. In *CVPR*, 2022. 4

[59] David Svitov, Dmitrii Gudkov, Renat Bashirov, and Victor Lempitsky. Dinar: Diffusion inpainting of neural textures for one-shot human avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7062–7072, 2023. 2

[60] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Trans. Graph.*, 38(4), jul 2019. 1, 2, 3, 5, 7, 13

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

[62] Shaofei Wang, Marko Mihajlovic, Qianli Ma, Andreas Geiger, and Siyu Tang. Metaavatar: Learning animatable clothed human models from few depth images. *Advances in Neural Information Processing Systems*, 34:2810–2822, 2021. 1

[63] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 5

[64] 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. 1

[65] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 7

[66] 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*, pages 16210–16220, 2022. 2

[67] Donglai Xiang, Timur Bagautdinov, Tuur Stuyck, Fabian Prada, Javier Romero, Weipeng Xu, Shunsuke Saito, Jingfan Guo, Breannan Smith, Takaaki Shiratori, et al. Dressing avatars: Deep photorealistic appearance for physically simulated clothing. *arXiv preprint arXiv:2206.15470*, 2022. 1, 3

[68] Jae Shin Yoon, Duygu Ceylan, Tuanfeng Y Wang, Jingwan Lu, Jimei Yang, Zhixin Shu, and Hyun Soo Park. Learning motion-dependent appearance for high-fidelity rendering of dynamic humans from a single camera. In *CVPR*, pages 3407–3417, 2022. 1, 3

[69] Meng Zhang, Tuanfeng Y Wang, Duygu Ceylan, and Niloy J Mitra. Dynamic neural garments. *ACM TOG*, 40(6):1–15, 2021. 1, 3

[70] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 7

[71] Hao Zhao, Jinsong Zhang, Yu-Kun Lai, Zerong Zheng, Yingdi Xie, Yebin Liu, and Kun Li. High-fidelity human avatars from a single rgb camera. In *CVPR*, pages 15904–15913, 2022. 2