Drivable Volumetric Avatars using Texel-Aligned Features

EDOARDO REMELLI, Meta Reality Labs, Switzerland
TIMUR BAGAUTDINOV, Meta Reality Labs, USA
SHUNSUKE SAITO, Meta Reality Labs, USA
TOMAS SIMON, Meta Reality Labs, USA
CHENGLEI WU, Meta Reality Labs, USA
SHIH-EN WEI, Meta Reality Labs, USA
KAIWEN GUO, Meta Reality Labs, USA
ZHE CAO, Meta Reality Labs, USA
FABIAN PRADA, Meta Reality Labs, USA
JASON SARAGIH, Meta Reality Labs, USA
YASER SHEIKH, Meta Reality Labs, USA

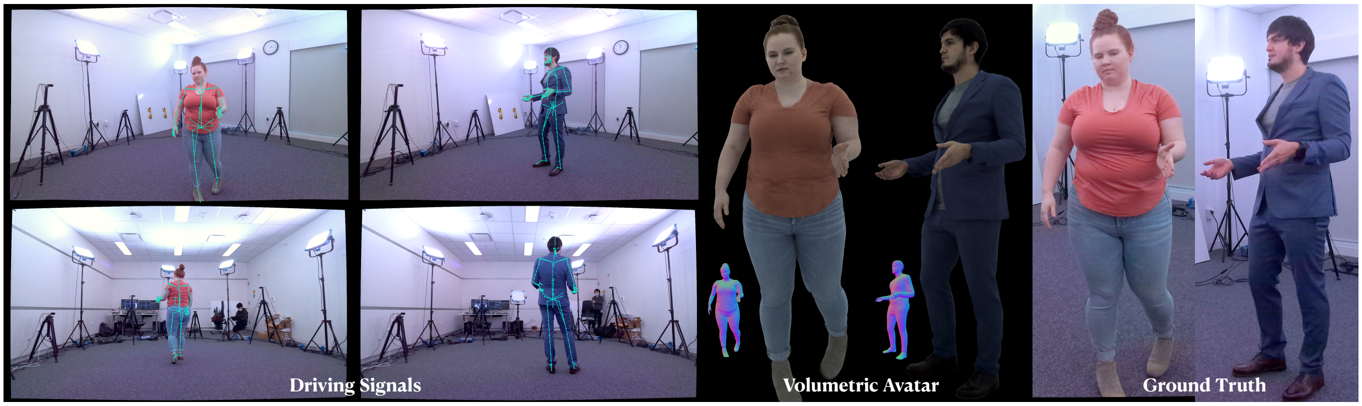

Fig. 1. We propose a drivable volumetric model for full-body avatars, which relies on texel-aligned features to be fully faithful to the driving signal.

Photorealistic telepresence requires both high-fidelity body modeling *and* faithful driving to enable dynamically synthesized appearance that is indistinguishable from reality. In this work, we propose an end-to-end framework that addresses two core challenges in modeling and driving full-body avatars of real people. One challenge is driving an avatar while staying faithful to details and dynamics that cannot be captured by a global low-dimensional parameterization such as body pose. Our approach supports driving of clothed avatars with wrinkles and motion that a real driving performer exhibits beyond the training corpus. Unlike existing global state representations or non-parametric screen-space approaches, we introduce texel-aligned features—a localised representation which can leverage both the structural prior of a skeleton-based parametric model and observed sparse image signals at the same time. Another challenge is modeling a temporally coherent clothed avatar, which typically requires precise surface tracking. To circumvent this, we propose a novel volumetric avatar representation by extending mixtures of volumetric primitives to articulated objects. By explicitly incorporating articulation, our approach naturally generalizes to unseen poses. We also introduce a localized viewpoint conditioning, which leads to a large improvement in generalization of view-dependent appearance. The proposed volumetric representation does not require high-quality mesh tracking as a prerequisite and brings significant quality improvements compared to mesh-based counterparts. In our experiments, we carefully examine our design choices and demonstrate the efficacy of our approach, outperforming the state-of-the-art methods on challenging driving scenarios.

CCS Concepts: • **Computing methodologies** → **Animation**.

Additional Key Words and Phrases: full-body avatar, volumetric representations, neural rendering

ACM Reference Format:
Edoardo Remelli, Timur Bagautdinov, Shunsuke Saito, Tomas Simon, Chenglei Wu, Shih-En Wei, Kaiwen Guo, Zhe Cao, Fabian Prada, Jason Saragih, and Yaser Sheikh. 2022. Drivable Volumetric Avatars using Texel-Aligned Features. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings (SIGGRAPH '22 Conference Proceedings), August 7–11, 2022, Vancouver, BC, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3528233.3530740

1 INTRODUCTION

Augmented reality (AR) and virtual reality (VR) have the potential to become major computing platforms, enabling people to interact with each other in ever more immersive ways across space and time.

SIGGRAPH '22 Conference Proceedings, August 7–11, 2022, Vancouver, BC, Canada

Among these possibilities, authentic social telepresence aims at lifelike presence in AR and VR which is indistinguishable from reality. This imposes a fundamental requirement for techniques to faithfully teleport every possible detail expressed by humans in reality.

One promising path to achieve this is to rely on a photorealistic animatable model, often obtained with an elaborate capture system, which essentially acts as a strong data-driven prior [Bagautdinov et al. 2021; Lombardi et al. 2018; Xiang et al. 2021]. Although these methods are capable of producing realistically-looking free-viewpoint renders, and are robust to occlusions and driving signal incompleteness, these methods do not fully exploit available inputs. In practice, such methods typically map dense sensory inputs to sparse driving signals, such as body pose or low-dimensional embeddings. Therefore, a large proportion of detailed observations about the subject are effectively thrown away, resulting in the need to re-hallucinate these details in the final render. This creates a clear fidelity gap between teleportation and reality, resulting in a loss in quality of the conveyed social cues. One of the reasons why relying exclusively on such model-based methods is insufficient lies in the fact that it is non-trivial to design a driving representation which is simultaneously expressive and relatively agnostic to the capture setup to ensure generalization in novel conditions.

An alternative path to building telepresence systems showing promise, is to rely on model-free methods which combine classical geometry reconstruction methods with image-space processing, either with ad-hoc image fusion [Lawrence et al. 2021] or neural re-rendering [Martin-Brualla et al. 2018; Shysheya et al. 2019]. Such methods are able to better exploit available inputs, but usually require highly specialized hardware for high-fidelity real-time sensing and would be limited in handling occlusions and incomplete data, either exhibiting limited quality rendering of novel viewpoints [Martin-Brualla et al. 2018] or requiring high viewpoint coverage by using multiple highly-specialized sensors [Lawrence et al. 2021]. Moreover, due to the challenges in real-time sensing, artifacts may occur around occlusion boundaries and body parts with limited resolution, like fingers or hair.

In this work, our goal is to design a method that effectively combines the expressiveness of model-free methods with the robustness of model-based neural rendering. The key idea is a localized state representation, which we call *texel-aligned features*, that provides the model with a dense conditioning signal while still relying on a data-driven neural rendering model. Dense conditioning allows us to maximize the amount of information extracted from the driving signals, while the data-driven model acts as a strong prior that allows the model to perform well even in scenarios with impoverished sensory input. Additionally, our approach uses a hybrid volumetric representation specifically tailored to modeling human bodies, which exhibits both good generalization to novel poses and produces high-quality free-viewpoint renders. This is in contrast to image-space neural rendering methods [Martin-Brualla et al. 2018] and mesh-based [Bagautdinov et al. 2021] methods, which either lead to poor generalization on novel views, or are not capable of modeling complex geometries with varying topology, which are abundant in clothed dynamic humans.

In our experiments, we demonstrate the effectiveness of such hybrid representations for full-bodies over the state-of-the-art. We also showcase the efficacy of our method by building a complete one-way telepresence system, which allows a person to be virtually teleported using only a few commodity sensors.

In summary, our contributions are:

- We introduce Drivable Volumetric Avatars (DVA): a novel neural representation for animation and free-viewpoint rendering of personalized human avatars.
- We propose texel-aligned features for DVA: a dense conditioning method that leads to better expressiveness and better generalization to novel viewpoints for unseen poses.
- We introduce a novel virtual teleportation system that uses DVA for one-way photorealistic telepresence.

Sample implementation will be made publicly available[1].

2 RELATED WORK

We first discuss existing representations for modeling dynamic human appearance and geometry. We then review existing telepresence systems, with a focus on how these systems exploit available driving signals, typically trading off robustness for fidelity.

Mesh-Based Avatars. Textured meshes have been widely used to represent human geometry and appearance for efficient rendering with modern graphics hardware. Parametric human body models can be learned from thousands of scans by deforming a template mesh [Anguelov et al. 2005; Hasler et al. 2009; Loper et al. 2015]. These approaches primarily focus on the geometry of minimally clothed human bodies. Recent works also model clothing shape variation [Bhatnagar et al. 2019; Ma et al. 2020]. Template meshes are also utilized to model the shape and appearance of clothed humans from video inputs [Alldieck et al. 2018] or a single image [Alldieck et al. 2019; Weng et al. 2019]. In particular, approaches [Grigorev et al. 2019; Lazova et al. 2019] are highly related to our work, as they leverage warping pixels to UV maps to perform novel view synthesis of clothed humans.

However, these methods model only static geometry and appearance, failing to produce high-fidelity drivable avatars for novel poses. Recently [Bagautdinov et al. 2021] proposed a method to model high-fidelity drivable avatars from a multi-view capture system by decoding dynamic geometry and appearance from disentangled driving signals. [Xiang et al. 2021] extends this representation by modeling clothing explicitly as a separate mesh layer, similarly to ClothCap [Pons-Moll et al. 2017], recovering sharper clothing boundaries.

Despite providing an efficient and effective way to represent dynamic humans, these mesh-based approaches require accurate tracking of the underlying geometry of avatars as a preprocessing step, which significantly limits supported clothing types. LiveCap [Habermann et al. 2019] and its follow-up learning-based method [Habermann et al. 2021] simplify the tracking requirement by leveraging silhouette constraints while achieving real-time performance. However, at the cost of simplification, the fidelity of the resulting avatars are not on par with the aforementioned approaches that rely on

[1] https://github.com/facebookresearch/dva

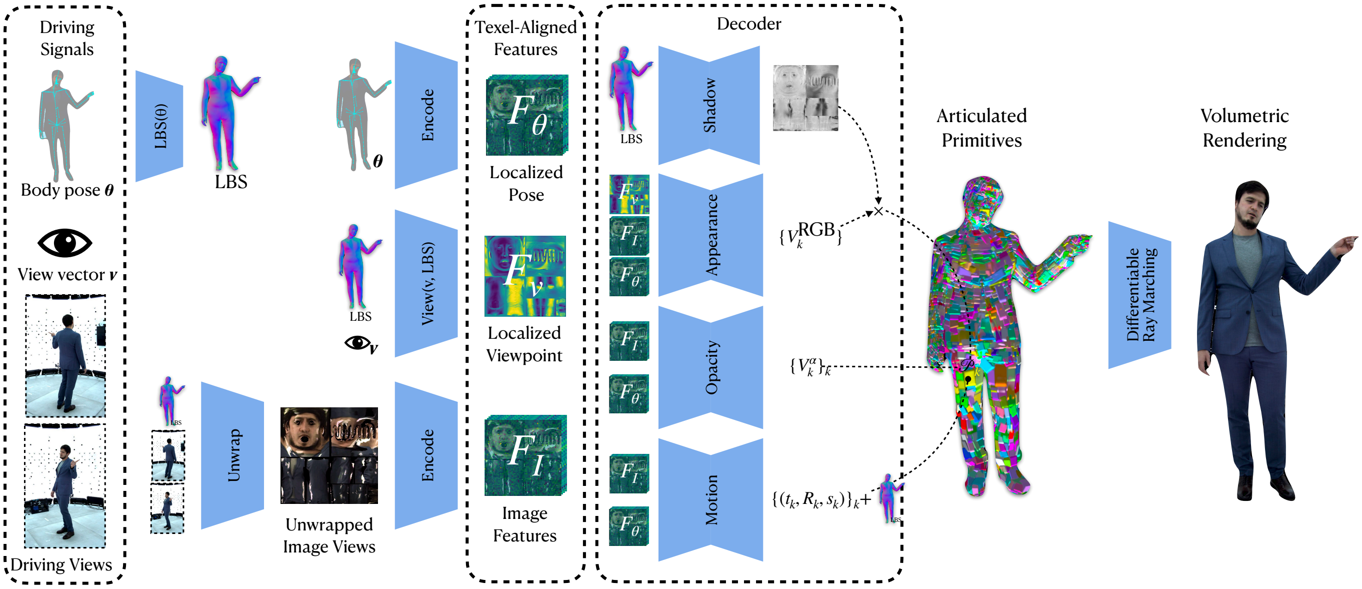

Fig. 2. **General overview of the architecture.** The core of our full-body model is a encoder-decoder architecture, which takes as input raw images, body pose, facial expression and view direction, and outputs a mixture of volumetric primitives. These are ray marched through to produce a full-body avatar.

accurate tracking. In contrast, our approach, is based on a more flexible volumetric representation, simplifying the tracking prerequisites while further improving the fidelity.

Volumetric Avatars. Recently volumetric representations have been shown effective for modeling 3D humans from a single image [Huang et al. 2020; Li et al. 2020a; Saito et al. 2019, 2020; Zheng et al. 2021], RGBD inputs [Li et al. 2020b; Yu et al. 2021], 3D scans [Bhatnagar et al. 2020; Chibane et al. 2020; Palafox et al. 2021; Saito et al. 2021; Tiwari et al. 2021], or multi-view captures [Liu et al. 2021; Lombardi et al. 2019, 2021; Peng et al. 2021a,b; Su et al. 2021].

PIFu [Saito et al. 2019] and its follow up works [Huang et al. 2020; Li et al. 2020a; Saito et al. 2020; Zheng et al. 2021] learn occupancy and texture fields given pixel-aligned image features and 3D coordinates. Extending it to RGBD inputs also enables robust avatar creation [Li et al. 2020b; Yu et al. 2021]. Parametric bodies such as SMPL model [Loper et al. 2015] are used to warp image-features to a canonical T-pose for modeling animatable avatars from a single image [He et al. 2021; Huang et al. 2020]. IFNet [Chibane et al. 2020] infers implicit functions from partial point clouds by leveraging multi-level features, which is later extended to multi-body parts [Bhatnagar et al. 2020]. Since volumetric representations support varying topology, modeling animatable clothed avatars is now possible without explicit surface registration [Palafox et al. 2021; Saito et al. 2021; Tiwari et al. 2021]. While some of these approaches model pose-dependent body geometry, the appearance is either ignored or not photo-realistic.

Given multi-view video sequences, Neural Volumes [Lombardi et al. 2019] models volumetric human heads by decoding radiance in voxels, whose efficiency and fidelity are further improved by Mixture of Volumetric Primitives (MVP) [Lombardi et al. 2021]. Similarly, [Ma et al. 2021a,b] introduce a collection of articulated primitives to model clothed humans, primarily focusing on geometry. Neural Body [Peng et al. 2021b] applies differentiable volumetric rendering [Mildenhall et al. 2020] to model articulated human geometry and appearance by diffusing per-vertex latent codes on a SMPL model via sparse 3D convolutions. Neural Actor [Liu et al. 2021] instead projects 3D coordinates on the closest SMPL surface to regress radiance fields. A-NeRF [Su et al. 2021] and [Peng et al. 2021a] leverages spatial transformations provided by joint articulation for better generalization with unseen poses. While the volumetric rendering approaches do not require precise tracking of the surface, they neither run in real-time nor model stochastic nature of clothing deformations as avatars are driven by only pose parameters. In contrast, our volumetric avatar runs in real-time, and supports more fine-grained control of the reconstructed avatars such as wrinkles and clothing dynamics.

Drivable Telepresence Systems. Drivable telepresence systems can be divided into two categories; Non-parametric solutions are holistically modeling a scene and faithfully transmit observed signals as it is [Lawrence et al. 2021; Martin-Brualla et al. 2018; Orts-Escolano et al. 2016]. The other solution is to leverage a category-specific parametric model for driving [Bagautdinov et al. 2021; Lombardi et al. 2018; Xiang et al. 2021].

Non-parametric telepresence systems, including Holoportation [Orts-Escolano et al. 2016], and Project Starline [Lawrence et al. 2021], leverage classical 3D reconstruction techniques and image-space post-processing for real-time rendering of the scene. While these approaches do not require per-user training, its fidelity is bounded by the input signals. Thus, occluded regions remain void. To alleviate this limitation, LookingGood [Martin-Brualla et al. 2018] proposes a system to augment real-time performance capture systems with 2D neural rendering. A neural network takes imperfect rendering of captured geometry and appearance, and jointly performs completion, super-resolution, and denoising in real-time. However, we argue that such a image-based solution lacks strong structural prior of human avatar, and results in sub-optimal rendering quality when input views are sparse (i.e., temporal flickering and blur in impainted regions). Such a 2D neural rendering technique has been recently applied to human body to better incorporate human prior [Prokudin et al. 2021; Raj et al. 2021; Shysheya et al. 2019]. However, these approaches rely exclusively on pose for driving the avatars, and do not capture remaining information such as clothing deformations and dynamics.

On the contrary, parametric approaches build strong subject-specific priors for face [Lombardi et al. 2018] and body [Bagautdinov et al. 2021; Habermann et al. 2021; Xiang et al. 2021] by training on a large corpus of multi-view data. While the advantage of these approaches lies in the robustness to limited input signals (e.g., driving from images from VR headsets [Wei et al. 2019]), high compression of drivers' state into low-dimensional driving signal often leads to lack of expressiveness for driving. In this work, our proposed texel-aligned feature representation allows us to leverage strong structural prior provided by a parametric model while recovering fine-grained details observed from sparse input images as in non-parametric systems.

3 METHOD

3.1 Overview

Our goal is to build a photorealistic personalized avatar of a human that is expressive and faithful to the driving signal, while also being robust to our driving setup with a sparse view inputs. An overview of our approach is provided in Fig. 2. Our model is an encoder-decoder architecture that takes as input a set of sparse multi-view images, body pose and a viewing direction, and produces a collection of volumetric primitives on a human body. The inferred volumetric primitives are then rendered with a differentiable ray marching to produce a photorealistic avatar. Input pose is used to produce coarse geometry articulated by Linear Blend Skinning [Kavan et al. 2008; Magnenat-Thalmann et al. 1989] (LBS). This provides the initial positions of the volumetric primitives. The underlying skinned model is also used to align all the available driving inputs onto *texel-aligned features*, which include localized pose, viewpoints, and image features. Those texel-aligned features are then decoded to a volumetric payload that captures high-resolution local geometry and appearance as well as dynamic correctives to the primitives' transformations. Similarly to [Bagautdinov et al. 2021], we use a shadow branch to capture non-local shading effects, which also operates in the texture space.

3.2 Background: Mixture of Volumetric Primitives

At the core of the MVP [Lombardi et al. 2021], the representation is a set of K dynamically moving volumetric primitives that jointly parameterize the color and opacity distribution of a modeled scene. This yields a scene representation that, unlike mesh-based ones, is not bound to a fixed topology and, compared to regular volumetric grids [Lombardi et al. 2019], is memory efficient and fast to render, allowing for real-time rendering of high-resolution dynamic scenes.

In practice, each primitive $\mathcal{P}_k = \{t_k, R_k, s_k, V_k^{\mathrm{RGB}}, V_k^{\alpha}\}$ is parameterized by a position $t_k \in \mathbb{R}^3$ in 3D space, an orientation $R_k \in \mathrm{SO}(3)$, a per-axis scale factor $s_k \in \mathbb{R}^3$, an appearance (RGB) payload $V_k^{\mathrm{rgb}} \in \mathbb{R}^{3\times S\times S\times S}$, and opacity $V_k^{\alpha} \in \mathbb{R}^{S\times S\times S}$, where S is the number of voxels along each spatial dimension.

The synthetic image is then obtained through differentiable ray marching, using a cumulative volumetric rendering scheme of [Lombardi et al. 2019]. More specifically, given a pixel p and corresponding ray $\mathbf{r}_p(t) = \mathbf{o} + t\mathbf{d}_p$ we compute its color I_p^{rgb} as

$$I_p^{\mathrm{rgb}} = \int_{t_{\min}}^{t_{\max}} V^{\mathrm{rgb}}(\mathbf{r}_p(t)) \frac{dT(t)}{dt} dt \,, \tag{1}$$

$$T(t) = \int_{t_{\min}}^{t_{\max}} V^{\alpha}(\mathbf{r}_p(t)) dt \,, \tag{2}$$

where $V^{\mathrm{RGB}}, V^{\alpha}$ denote global color and opacity fields and are computed by trilinearly interpolating each primitive hit by the ray.

3.3 Articulated Primitives

[Lombardi et al. 2021] introduces a model for human faces, where volumetric primitives are loosely attached to a guide mesh directly regressed by an MLP. Unlike faces, however, bodies undergo large rigid motions that are hard to handle robustly with position-based regression. In order to generalize better to articulated motion, we propose to attach primitives to the output mesh of Linear Blend Skinning model:

$$\mathcal{M}_{\boldsymbol{\theta}} = \mathrm{LBS}(\boldsymbol{\theta}, \mathcal{M}) \,, \tag{3}$$

where θ denotes human pose, $\mathcal{M}$ a template mesh in canonical pose, and $\mathcal{M}_{\theta}$ is the final mesh geometry after posing.

We then initialize primitive locations by uniformly sampling UV-space, mapping each primitive to the closest texel, and positioning it at the corresponding surface point $\hat{t}_k(\boldsymbol{\theta}) \in \mathcal{M}_{\boldsymbol{\theta}}$. In practice, we use $K = 4096$ primitives, and thus this procedure produces a $W{\times}W$, $W = 64$ grid, where each primitive is *aligned* to a specific *texel*. The orientation of the primitives is initialized based on the local tangent frame $\hat{R}_k(\boldsymbol{\theta})$ of the 3D surface point on the reposed mesh, and the scale $\hat{s}_k$ of each primitive is initialized based on the gradient of 3D rest shape w.r.t.the UV-coordinates at the corresponding grid point position.

Moreover, although the primitives are associated with the articulated mesh, in order to allow for larger variations in topology, they are allowed to deviate:

$$t_k = \delta t_k + \hat{t}_k(\boldsymbol{\theta}), \tag{4}$$

$$R_k = \delta R_k \cdot \hat{R}_k(\boldsymbol{\theta}), \tag{5}$$

$$s_k = \delta s_k + \hat{s}_k, \tag{6}$$

where $\delta t_k, \delta R_k, \delta s_k$ are primitive *correctives* that are produced by the motion branch of our decoder.

3.4 Texel-Aligned Features

In this section, we describe our novel dense representation which allows us to fully exploit available driving signals. As discussed earlier, our primitives are aligned into a $W \times W$ 2D grid, where each primitive is assigned to a specific texel on the UV-map. In order to condition each of the primitives only on the relevant information, we propose to use the same spatial prior to align all the input signals to the corresponding structure.

Body Pose. For body pose $\boldsymbol{\theta}$, we employ location-specific encodings similar to [Bagautdinov et al. 2021]: we use skinning weights to limit the spatial extent of each pose parameter, project the resulting masked features to the UV-space at the same $W \times W$ resolution, and then apply a dimensionality reducing projection, implemented as a 1x1 convolution, to get $F_{\boldsymbol{\theta}}$. Such localized representation helps limit overfitting, and reduces the tendency to learn spurious long-range correlations which might be present in the training data as shown in prior works [Bagautdinov et al. 2021; Saito et al. 2021].

Images. Conditioning the model only on pose is insufficient for faithful driving, as it does not contain all the information required to explain the entire appearance of a clothed human in motion, such as stochastic clothing state and dynamics. To this end, we propose to use available image evidence, by projecting it to a common texture space. Namely, for each available input view, we back-project image pixels corresponding to all visible vertices of our posed LBS template to UV-domain. Once all visual evidence has been mapped to a common UV-domain, we average it across all the available views to get a multi-view texture. We then compress the resulting high-resolution texture to the same resolution as $F_{\boldsymbol{\theta}}$ with a convolutional encoder to obtain F_{I}. Note that, unlike the existing works [Bagautdinov et al. 2021; Lombardi et al. 2018], which assume all the remaining information is encoded into a global low-dimensional code, our representation preserves spatial structure of the signal and is significantly more expressive, thus allowing to better capture deformations present in the input signals (see Fig.4).

Viewpoint. The existing work on modeling view-dependent appearance of human body [Bagautdinov et al. 2021; Liu et al. 2021; Peng et al. 2021a,b] represents viewpoint globally for the entire body, as a relative camera position w.r.t.the root joint. However, this representation is not explicitly taking into account articulation, and the model is forced to learn complex interactions between pose-dependent deformations and viewpoint from limited data, which leads to overfitting (see Fig. 3). To address this, we encode the camera position in the *local* coordinate frame of each primitive. Specifically, given a view-direction $\mathbf{v} \in \mathbb{R}^3$ and the posed template mesh $\mathcal{M}_\theta$, we compute per-triangle normals $\mathbf{n}_t$ and use them to express camera coordinates relatively to the local tangent plane of each triangle as

$$v_t = \mathbf{v} \cdot \mathbf{n}_t. \tag{7}$$

Then, we warp this quantity to UV space, and sample it at $W \times W$ resolution to get texel-aligned viewpoint features F_{v}.

3.5 Architecture and Training Details

Given texel-aligned features, our decoder produces the payload of the volumetric primitives (Fig. 2). In practice, we employ three independent branches, each being a sequence of 2D transposed convolutions with untied biases that preserve spatial alignment. The motion branch is conditioned on $(F_{\boldsymbol{\theta}}, F_{\mathrm{I}})$, and produces transformation correctives $\{(\delta t_k, \delta R_k, \delta s_k)\}_k \in \mathbb{R}^{9\times W\times W}$, which are then applied to the initial locations of articulated primitives. The opacity branch is conditioned on $(F_{\boldsymbol{\theta}}, F_{\mathrm{I}})$ and produces a slab $\{V_k^\alpha\}_k \in \mathbb{R}^{S\times W\cdot S\times W\cdot S}$, where $S = 16$ is the number of voxels along a spatial dimension. The appearance branch is conditioned on $(F_{\boldsymbol{\theta}}, F_{\mathrm{I}}, F_{\mathrm{v}})$, and produces a slab $\{V_k^{\mathrm{rgb}}\}_k \in \mathbb{R}^{3\times S\times W\cdot S\times W\cdot S}$. Additionally, to capture long-range pose-dependent effects, we employ a shadow branch in [Bagautdinov et al. 2021] by replacing the output channel of the last convolution layer to match with the V_k^{rgb} for multiplication.

We use the following composite loss to train all our models:

$$\mathcal{L} = \lambda_{\mathrm{rgb}}\mathcal{L}_{\mathrm{rgb}} + \lambda_{\mathrm{vgg}}\mathcal{L}_{\mathrm{vgg}} + \lambda_{\mathrm{m}}\mathcal{L}_{\mathrm{m}} + \lambda_{\mathrm{vol}}\mathcal{L}_{\mathrm{vol}}\ , \tag{8}$$

where $\mathcal{L}_{\mathrm{rgb}}$ is the MSE image loss, $\mathcal{L}_{\mathrm{vgg}}$ is the perceptual VGG-loss, $\mathcal{L}_{\mathrm{m}}$ is the MAE segmentation mask loss, and $\mathcal{L}_{\mathrm{vol}}$ is the volume prior loss [Lombardi et al. 2021] that encourages primitives to be as small as possible. Empirically, we found that it is important to train the model in two stages to ensure robustness with respect to the quality of tracking and LBS model. Namely, for the first $N = 1000$ iterations we condition our model on all available training views (for our data 160 cameras), and then continue training on a sparse signal computed from 3 randomly sampled view. Intuitively, this helps our model pick up useful signals by gradually shifting from easy samples to harder ones in the spirit of curriculum learning [Bengio et al. 2009].

4 EXPERIMENTS

In this section, we report our experimental findings, ablate different components of our method (DVA), and showcase a teleportation system that uses DVA to create a one-way photorealistic telepresence experience.

4.1 Datasets

We report most of our results on data acquired with a setup similar to [Bagautdinov et al. 2021; Xiang et al. 2021]: a multi-view dome-shaped rig with 160 high-resolution (4K) synchronized cameras. We collect data for three different identities, including one with challenging multi-layer clothing (a suit), run LBS-tracking pipeline to obtain ground truth poses, and then use roughly 1000 frames in various poses for training our models. Additionally, we evaluate the performance of our articulated volumetric representation for human bodies on a public dataset ZJU-MoCAP [Peng et al. 2021b][2].

4.2 Ablation Study

View Conditioning. In this experiment, we demonstrate the effectiveness of the localized view conditioning described in Section 3.4. In Fig. 3, we show qualitative performance of a version of our model trained with local view conditioning and the instance trained with the global one [Bagautdinov et al. 2021]. Our localized view conditioning leads to plausible view-dependent appearance with unseen poses, whereas the global view conditioning suffers from significant visual artifacts due to overfitting.

[2] No facial meshes were created for the individuals in ZJU-MoCAP, and the dataset was not used to identify individuals

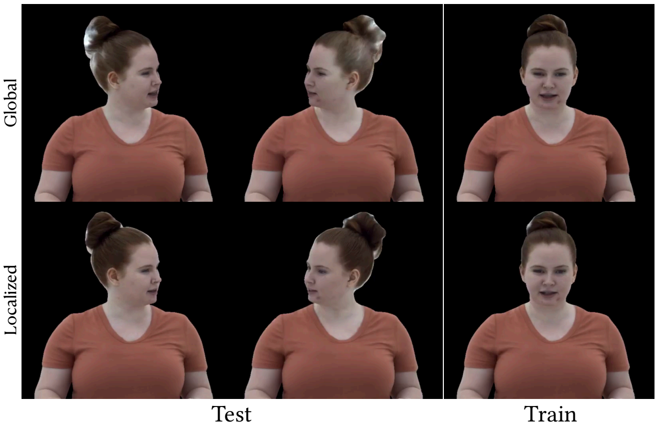

Fig. 3. **Effects of view conditioning.** Localized view conditioning leads to better generalization on unseen combinations of poses and viewpoints.

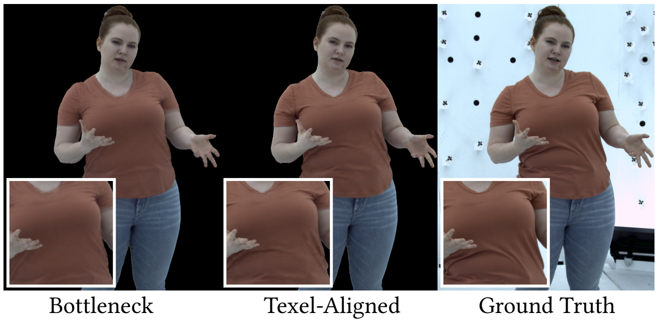

Fig. 4. **Effects of texel-aligned features.** Expressive texel-aligned features allow our model to generalize better to challenging unseen clothing states.

Texel-Aligned Features. In Fig. 4, we provide qualitative comparison of two different instances of our method: one that uses texel-aligned features, and one that relies on a bottleneck representation akin to [Bagautdinov et al. 2021]. The instance with texel-aligned features demonstrates significantly better preservation of high-frequency details with respect to the ground truth.

4.3 Novel View Synthesis

In order to evaluate the effectiveness of our articulated volumetric representation with respect to existing methods, we provide quantitative and qualitative comparisons on ZJU-MoCap [Peng et al. 2021b] dataset. The goal of this challenging benchmark is to produce a photorealistic novel view synthesis (NVS) of a clothed human in motion, while training only from 4 views.

Table 1. Quantitative results (PSNR) for NVS on ZJU-MoCap.

Method	S386	S387
FBCA	32.123	27.886
NeuralBody	33.196	28.640
OURS	**35.414**	**30.512**

We compare to Full-Body Codec Avatars (FBCA) [Bagautdinov et al. 2021] and Neural Body [Peng et al. 2021b], which represent the state-of-the-art among respectively mesh-based and volumetric approaches. In Tab. 1, we provide quantitative results for novel view synthesis in terms of PSNR for two subjects, which suggest that our method provides significant improvements over baselines. In Fig. 5, we provide a qualitative comparison; our method produces sharper reconstructions and less artifacts than both of the baselines. Interestingly, the mesh-based FBCA is performing significantly worse in settings without direct mesh supervision by precise multi-view stereo and tracking (the only source of geometry supervision in ZJU-MoCap dataset is silhouette and image losses through differentiable rendering). In constrast, our volumetric approach is able to learn more accurate underlying geometry with only image-based supervision due to the flexibility of volumetric representation. Please refer to the supplemental video for more detailed visual comparison.

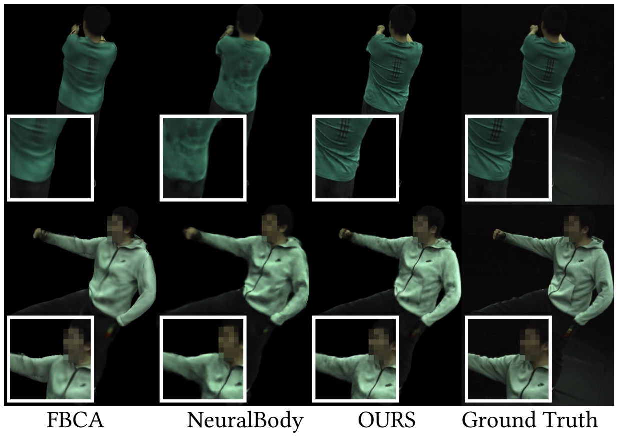

Fig. 5. **Novel View Synthesis.** We compare our method to state-of-the-art for NVS on ZJU-MoCap. Despite not being explicitly tailored to be trained with sparse supervision, our method outperforms competitors. Real faces and their reconstructions are blurred for anonymity.

4.4 Driving Results

We compare our approach to two different kinds of drivable telepresence systems: a mesh-based model (FBCA), and image-space model (LookingGood), both trained on our high-quality multi-view captures. Note that, in practice we use our own re-implementation of LookingGood, which uses LBS tracking instead of raw depth maps.

Table 2. Quantitative results (PSNR) on unseen motion. Please refer to supplementary video for more results.

Method	Views	Test view		
		`front`	`back`	avg
FBCA	None	30.571	30.656	30.613
LookingGood	2	33.442	26.059	29.750
	3	33.256	32.453	32.854
OURS	2	33.615	33.203	33.409
	3	**33.617**	**33.838**	**33.728**

Results of quantitative evaluation are provided in Table 2. Models are evaluated on two different views, `front`-facing and `back`-facing. To evaluate robustness to the sparsity of input views, we consider two different settings for the approaches that utilize input images as driving signal: one where we provide only conditioning from 2 front facing cameras, and a less challenging one where we provide 3 uniformly sampled conditioning views. Our model outperforms both of the baselines, and is more robust to missing information compared to LookingGood, in particular on settings with severe sensory deprivation. Qualitative results are provided in Fig. 6. We provide additional comparisons on dynamic sequences in supplementary video.

4.5 Teleportation

We also demonstrate the versatility of our method and show that our reconstructed avatars can be driven outside the capture system used to generate training data without losing details. Our setup consists of 8 synchronized and calibrated Microsoft Azure Kinects cameras [3], uniformly placed in a circle of 4.5-meter diameter. To obtain body poses (LBS parameters), we fit a pre-built personalized LBS body model to a sequence of detected and triangularized keypoints from RGB images [Wei et al. 2016], as well as meshes obtained by fusing multiview point clouds [Yu et al. 2021]. To obtain texel-aligned features, we simply apply texture unwrapping as in our data processing for the capture dome. Even though these driving signals are obtained from unseen sequences under different sensor modality and pre-processing, we can still faithfully animate our avatars. Figure 1 shows that the animated avatars preserve local details such as wrinkles on the clothes without noticeable artifacts in appearance. For more results, please refer to the supplementary video.

5 CONCLUSION

We introduced Drivable Volumetric Avatars, a novel method for building expressive fully articulated avatars and faithfully driving it from sparse view inputs. Our approach combines the robustness of parametric models by incorporating a strong articulated volumetric prior, and the expressiveness of non-parametric models by leveraging texel-aligned features. We demonstrated the efficacy of our method on novel view synthesis and driving scenarios, and showcased a one-way teleportation system based on our approach to create a photorealistic telepresence experience. Some of the main limitations of our work originate from our reliance on LBS tracking: our model still requires cumbersome skeleton tracking as a pre-processing step, and cannot handle very loose clothing that significantly deviates from the guide LBS mesh. A potential avenue for future work is extending the model to multi-identity settings, multiple outfits, and driving from a head-mounted capture device for a two-way telepresence system.

[3] https://azure.microsoft.com/en-us/services/kinect-dk/

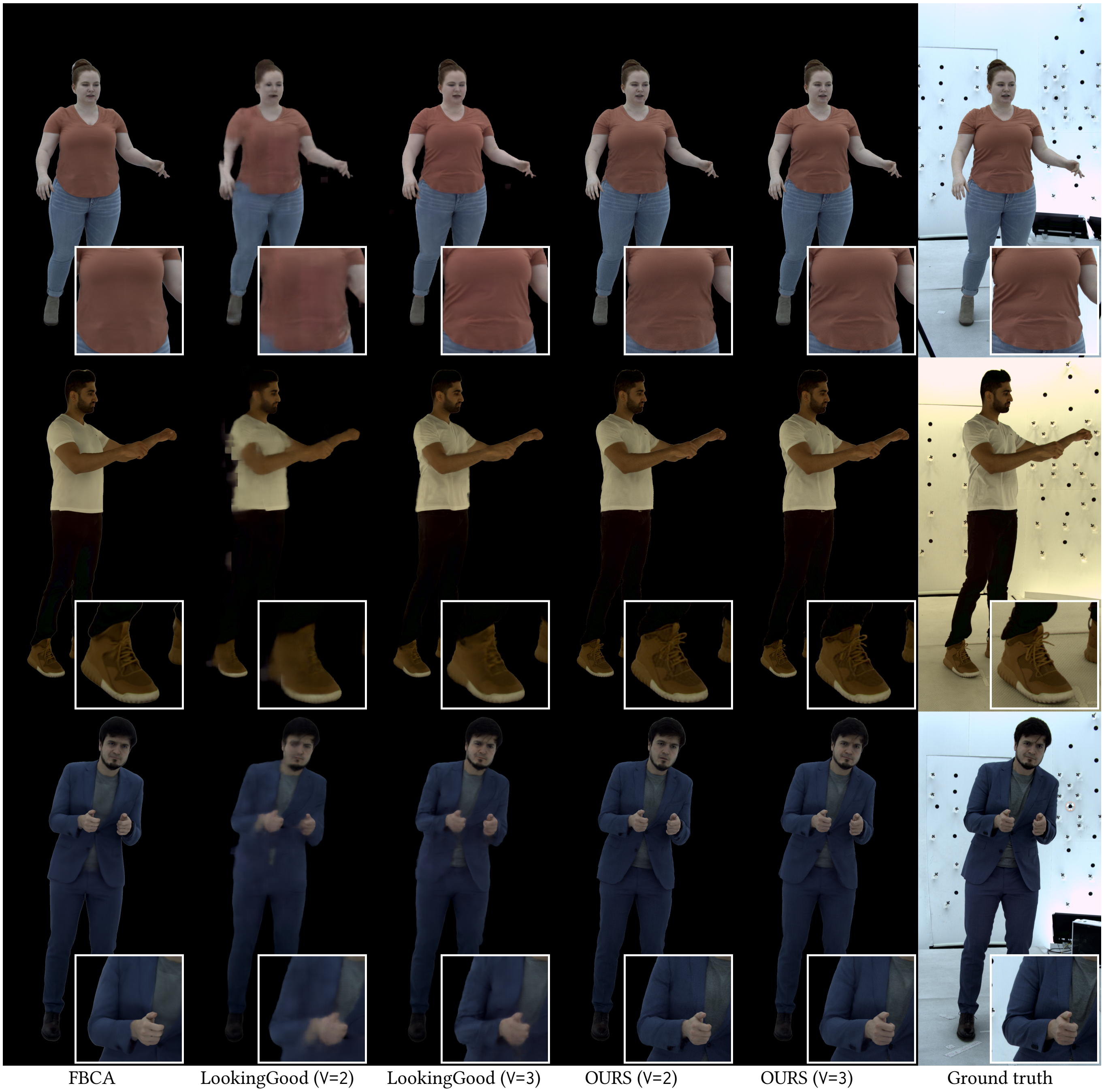

Fig. 6. **Qualitative results: Driving.** We compare our method to state-of-the-art approaches for drivable avatars on unseen sequences from our dataset. Best seen in supplemental video. V is the number of view inputs.

REFERENCES

Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. 2018. Video Based Reconstruction of 3D People Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. 2019. Tex2shape: Detailed full human body geometry from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2293–2303.

Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. 2005. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*. 408–416.

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

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning.. In *ICML (ACM International Conference Proceeding Series, Vol. 382)*, Andrea Pohoreckyj Danyluk, Léon Bottou, and Michael L. Littman (Eds.). ACM, 41–48. http://dblp.uni-trier.de/db/conf/icml/icml2009.html#BengioLCW09

Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. 2020. Combining Implicit Function Learning and Parametric Models for 3D Human Reconstruction. In *European Conference on Computer Vision (ECCV)*. Springer.

Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. 2019. Multi-garment net: Learning to dress 3d people from images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5420–5430.

Julian Chibane, Thiemo Alldieck, and Gerard Pons-Moll. 2020. Implicit Functions in Feature Space for 3D Shape Reconstruction and Completion. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Artur Grigorev, Artem Sevastopolsky, Alexander Vakhitov, and Victor Lempitsky. 2019. Coordinate-based texture inpainting for pose-guided human image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12135–12144.

Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. 2021. Real-time Deep Dynamic Characters. *arXiv preprint arXiv:2105.01794* (2021).

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

Nils Hasler, Carsten Stoll, Martin Sunkel, Bodo Rosenhahn, and H-P Seidel. 2009. A statistical model of human pose and body shape. In *Computer graphics forum*, Vol. 28. Wiley Online Library, 337–346.

Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. 2021. ARCH++: Animation-Ready Clothed Human Reconstruction Revisited. *ArXiv* abs/2108.07845 (2021).

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

Ladislav Kavan, Steven Collins, Jiří Žára, and Carol O'Sullivan. 2008. Geometric Skinning with Approximate Dual Quaternion Blending. *ACM Trans. Graph.* 27, 4, Article 105 (Nov. 2008), 23 pages. https://doi.org/10.1145/1409625.1409627

Jason Lawrence, Dan B Goldman, Supreeth Achar, Gregory Major Blascovich, Joseph G Desloge, Tommy Fortes, Eric M Gomez, Sascha Häberling, Hugues Hoppe, Andy Huibers, et al. 2021. Project Starline: A high-fidelity telepresence system. (2021).

Verica Lazova, Eldar Insafutdinov, and Gerard Pons-Moll. 2019. 360-degree textures of people in clothing from a single image. In *2019 International Conference on 3D Vision (3DV)*. IEEE, 643–653.

Ruilong Li, Yuliang Xiu, Shunsuke Saito, Zeng Huang, Kyle Olszewski, and Hao Li. 2020a. Monocular Real-Time Volumetric Performance Capture. *arXiv preprint arXiv:2007.13988* (2020).

Zhe Li, Tao Yu, Chuanyu Pan, Zerong Zheng, and Yebin Liu. 2020b. Robust 3D Self-portraits in Seconds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. http://arxiv.org/abs/2004.02460v1

Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. 2021. Neural Actor: Neural Free-view Synthesis of Human Actors with Pose Control. *arXiv preprint arXiv:2106.02019* (2021).

Stephen Lombardi, Jason Saragih, Tomas Simon, and Yaser Sheikh. 2018. Deep appearance models for face rendering. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–13.

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

Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. 2021. Mixture of volumetric primitives for efficient neural rendering. *arXiv preprint arXiv:2103.01954* (2021).

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

Qianli Ma, Shunsuke Saito, Jinlong Yang, Siyu Tang, and Michael J. Black. 2021a. SCALE: Modeling Clothed Humans with a Surface Codec of Articulated Local Elements. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Qianli Ma, Jinlong Yang, Anurag Ranjan, Sergi Pujades, Gerard Pons-Moll, Siyu Tang, and Michael J Black. 2020. Learning to dress 3d people in generative clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6469–6478.

Qianli Ma, Jinlong Yang, Siyu Tang, and Michael J. Black. 2021b. The Power of Points for Modeling Humans in Clothing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

N. Magnenat-Thalmann, R. Laperrière, and D. Thalmann. 1989. Joint-Dependent Local Deformations for Hand Animation and Object Grasping. In *Proceedings on Graphics Interface '88* (Edmonton, Alberta, Canada). Canadian Information Processing Society, CAN, 26–33.

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

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*. Springer, 405–421.

Sergio Orts-Escolano, Christoph Rhemann, Sean Fanello, Wayne Chang, Adarsh Kowdle, Yury Degtyarev, David Kim, Philip L Davidson, Sameh Khamis, Mingsong Dou, et al. 2016. Holoportation: Virtual 3d teleportation in real-time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. 741–754.

Pablo Palafox, Aljaž Božič, Justus Thies, Matthias Nießner, and Angela Dai. 2021. NPMs: Neural Parametric Models for 3D Deformable Shapes. (2021).

Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. 2021a. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14314–14323.

Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. 2021b. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9054–9063.

Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael Black. 2017. ClothCap: Seamless 4D Clothing Capture and Retargeting. *ACM Transactions on Graphics, (Proc. SIGGRAPH)* 36, 4 (2017). http://dx.doi.org/10.1145/3072959.3073711 Two first authors contributed equally.

Sergey Prokudin, Michael J. Black, and Javier Romero. 2021. SMPLpix: Neural Avatars from 3D Human Models. In *Winter Conference on Applications of Computer Vision (WACV)*. 1810–1819.

Amit Raj, Julian Tanke, James Hays, Minh Vo, Carsten Stoll, and Christoph Lassner. 2021. ANR: Articulated Neural Rendering for Virtual Avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3722–3731.

Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. 2019. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2304–2314.

Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. 2020. PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J. Black. 2021. SCANimate: Weakly Supervised Learning of Skinned Clothed Avatar Networks. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Aliaksandra Shysheya, Egor Zakharov, Kara-Ali Aliev, Renat Bashirov, Egor Burkov, Karim Iskakov, Aleksei Ivakhnenko, Yury Malkov, Igor Pasechnik, Dmitry Ulyanov, et al. 2019. Textured neural avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2387–2397.

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

Garvita Tiwari, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. 2021. Neural-GIF: Neural Generalized Implicit Functions for Animating People in Clothing. In *International Conference on Computer Vision (ICCV)*.

Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional pose machines. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 4724–4732.

Shih-En Wei, Jason Saragih, Tomas Simon, Adam W. Harley, Stephen Lombardi, Michal Perdoch, Alexander Hypes, Dawei Wang, Hernan Badino, and Yaser Sheikh. 2019. VR Facial Animation via Multiview Image Translation. *ACM Trans. Graph.* 38, 4, Article 67 (jul 2019), 16 pages. https://doi.org/10.1145/3306346.3323030

Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. 2019. Photo wake-up: 3d character animation from a single photo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5908–5917.

Donglai Xiang, Fabian Prada, Timur Bagautdinov, Weipeng Xu, Yuan Dong, He Wen, Jessica Hodgins, and Chenglei Wu. 2021. Modeling Clothing as a Separate Layer for an Animatable Human Avatar. *arXiv preprint arXiv:2106.14879* (2021).

Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. 2021. Function4D: Real-time Human Volumetric Capture from Very Sparse Consumer RGBD Sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 5746–5756.

Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. 2021. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).