Vid2Avatar-Pro: Authentic Avatar from Videos in the Wild via Universal Prior

Chen Guo*[1,2] Junxuan Li*[1] Yash Kant[1,3] Yaser Sheikh[1] Shunsuke Saito†[1] Chen Cao†[1]
[1]Codec Avatars Lab, Meta [2]ETH Zürich [3]University of Toronto

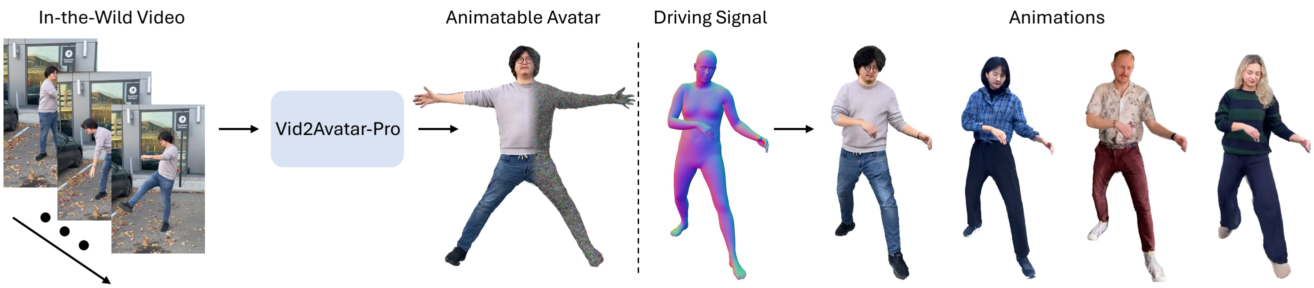

Figure 1. We present Vid2Avatar-Pro, a method to create photorealistic 3D human avatars from monocular in-the-wild videos via pre-trained universal prior. Our method can faithfully create high-fidelity human avatars from a single video and generate realistic animations.

Abstract

*We present Vid2Avatar-Pro, a method to create photorealistic and animatable 3D human avatars from monocular in-the-wild videos. Building a high-quality avatar that supports animation with diverse poses from a monocular video is challenging because the observation of pose diversity and view points is inherently limited. The lack of pose variations typically leads to poor generalization to novel poses, and avatars can easily overfit to limited input view points, producing artifacts and distortions from other views. In this work, we address these limitations by leveraging a universal prior model (UPM) learned from a large corpus of multi-view clothed human performance capture data. We build our representation on top of expressive 3D Gaussians with canonical front and back maps shared across identities. Once the UPM is learned to accurately reproduce the large-scale multi-view human images, we fine-tune the model with an in-the-wild video via inverse rendering to obtain a personalized photorealistic human avatar that can be faithfully animated to novel human motions and rendered from novel views. The experiments show that our approach based on the learned universal prior sets a new state-of-the-art in monocular avatar reconstruction by substantially outperforming existing approaches relying only on heuristic regularization or a shape prior of minimally clothed bodies (*e.g.*, SMPL) on publicly available datasets.*

*Equal contribution †Equal advisory

1. Introduction

Authentic digital humans are widely used for the synthesis of novel animations of real persons in games and movies. They are also expected to be an indispensable component for virtual communications with immersive display devices (AR/VR). However, creating such an authentic avatar typically requires expensive multi-view capture systems [2], limiting its availability to professional studios. If we can effortlessly build a high-quality avatar just from an in-the-wild monocular video, it unlocks myriads of applications not only for professionals but also for everybody.

Despite its promise, it remains non-trivial to create a high-quality avatar from in-the-wild videos that can be faithfully animated with diverse poses and efficiently rendered from arbitrary views. Monocular videos often do not cover the entire space of pose and view points required for test-time animation and rendering. The scarcity of pose variations leads to poor generalization to novel poses. Specifically, the avatar may exhibit artifacts and unnatural deformations when animated using out-of-distribution poses. Moreover, inverse rendering with limited view coverage is prone to overfitting, resulting in distortions and artifacts when rendered from unseen camera views.

Existing approaches [15, 22, 29, 52, 61, 78, 88] aim to overcome its limitations by incorporating statistical prior from a minimally-clothed human body model [50, 57] or geometric prior based on heuristics (*e.g.*, Laplacian regularization [52]). While these priors improve robustness under this ill-posed problem setting, there remains a clear quality gap from 3D human avatars reconstructed from high-quality studio data especially when animated with novel poses.

In this work, we argue that the core problem lies in the fact that the aforementioned priors are not built for clothed human avatar modeling. Parametric models such as SMPL [50] focus on minimally clothed bodies without appearance. Laplacian regularization uniformly penalizes deformations regardless of the underlying materials. To address this, we introduce Vid2Avatar-Pro, a method to create photorealistic 3D human avatars from monocular in-the-wild videos by leveraging universal prior directly learned from high-quality clothed human performance capture data.

Our universal prior model is based on 3D Gaussians [34] due to its efficiency and expressiveness in representing details. While existing universal models for face and hands [5, 10, 41] rely on shared UV parameterization, it is not suitable for clothed humans due to the diverse topologies of clothing. Inspired by [46], we propose, for the first time, front and back maps as a universal parametrization for clothed humans. More specifically, we warp static SDF-based reconstructions [15, 86] of all subjects to a canonical space by normalizing both poses and bone lengths, and obtain color and position maps from front and back views with orthogonal projection. This allows us to maximize the spatial alignment of identity conditioning data across identities while supporting diverse topologies of clothing. We feed the front and back maps to a U-Net [5, 41, 64] to predict pose-dependent Gaussian attributes in a pixel-aligned manner with the input maps. We train this universal prior model via inverse rendering with multi-view performance capture of a thousand of clothed humans. Once trained, we personalize the universal prior model to monocular in-the-wild videos. We first reconstruct the canonical textured template from the monocular observations using neural SDF [15, 86]. Given the conditioning data obtained from the canonical textured template, we inpaint the unseen texture regions using a diffusion model [12] adapted for inpainting. We then fine-tune the model via inverse rendering and update the network weights of the prior model to faithfully reconstruct person-specific details from the monocular observations.

The experiments show that our approach produces photorealistic avatars that can be animated beyond training pose distributions with faithful pose-dependent deformations and appearance changes (*cf*. Fig. 1). We also carefully validate our design choices via ablation studies. Furthermore, we compare our method with state-of-the-art approaches in novel view/pose synthesis using publicly available datasets, and demonstrate that Vid2Avatar-Pro outperforms them by a substantial margin both quantitatively and qualitatively.

In summary, our contributions are:

- A universal prior model of clothed humans directly learned from a thousand of high-quality dynamic performance capture data.
- The first universal prior model architecture designed to use spatially normalized front and back identity conditioning data, enabling efficient and scalable training of multi-identity clothed human avatars.
- A robust personalization pipeline to create photorealistic and animatable clothed human avatars from monocular in-the-wild videos, achieving new state-of-the-art performance.

2. Related Work

Avatar Reconstruction from Multi-View Videos. High-fidelity 3D avatar reconstruction has required calibrated multi-view systems [2, 9, 18, 23, 26, 44, 45, 48, 55, 59, 60, 62, 67, 69, 75, 79, 83, 87, 94, 95]. These approaches utilize neural rendering techniques (*e.g.*, [51, 86]) to learn an implicit canonical representation of clothed humans. More recently, researchers retain the human skeletal structure and substituting implicit neural rendering with 3DGS [34] to learn person-specific animatable avatars [32, 46, 54, 56, 92, 96]. Our front and back map representation is inspired by [21, 46, 56], specifically Li *et al*. [46]. Unlike UV parameterization which necessitates manual efforts [56] or only covers minimally-clothed bodies [21], our front and back parameterization can be obtained automatically and accommodates various topologies of clothed humans. Different from Li *et al*. [46], who use front and back position maps as pose features for single identity modeling, we propose, for the first time, to extend such parameterization to be universal across multiple identities. Compared to aforementioned methods that require a specialized capturing setup, our approach creates photorealistic and animatable avatars from just monocular RGB videos via learned universal prior.

Avatar Reconstruction from Monocular Video. Traditional mesh-based methods [14, 17, 50, 57, 63, 84] are limited to a fixed topology and resolution, and cannot represent fine-grained details like the human faces. Regression-based methods that learn to regress 3D human shape and texture from images have demonstrated compelling results [1, 13, 19, 20, 24, 25, 65, 66, 80–82, 93]. A major limitation of these methods is that the reconstructed avatars can only be driven using rigid skeleton movement without pose-dependent deformations, leading to unrealistic animation results. Fitting articulated implicit neural fields to monocular videos via inverse neural rendering has been demonstrated [15, 16, 27–30, 70, 71, 78, 85, 88]. Following this line of work, emerging methods fit articulated 3D Gaussians to the monocular videos by either directly optimizing the 3D Gaussian attributes [40, 43, 68, 74] or training neural networks to predict the attributes [21, 22, 37, 42, 49, 52, 61, 77]. GaussianAvatar [21] leverages a UV positional map of SMPL to generate pose-dependent effects. ExAvatar [52] combines with SMPL-X to achieve a whole-body avatar with face and hand control. These methods achieve improved rendering speed and quality. However, the reconstructed avatars still show blurriness and lack appearance details. Especially, they tend to generate uncanny animation results for poses that are out of the training distribution from the short video. In contrast, our method leverages a high-quality and large-scale corpus of dynamic human captures to train a universal prior model, which can be personalized to short video data and ensure superior generalization ability to novel poses.

Human-Centric Prior Model. Statistical body models [31, 50, 57] can hardly model human surface details such as clothing and facial features. Regression-based methods [6, 19, 20, 25, 65, 66, 80, 81] train networks to learn pixel-aligned human features. These models are trained on small-scale static 3D human scans, thus the reconstructed humans are often not animatable or can only show rigid and unrealistic animations. Generalizable human rendering methods are trained on small-scale multi-view images to synthesize human novel views from sparse camera inputs without supporting human controllability [7, 8, 38, 39, 72, 90, 91]. There are some promising prior models that recover detailed appearance and generate plausible animations for human hands [10, 11, 53] and faces [3–5, 41]. In this work, we propose a universal prior model for photorealistic and animatable clothed humans, learned from thousands of dynamic human performance captures with diverse identities, garment styles, view points and poses.

3. Method

We introduce Vid2Avatar-Pro, a method to create photorealistic 3D human avatars from monocular in-the-wild videos. Our method is illustrated in Fig. 2 and consists of two main steps: building universal prior model for clothed humans (Sec. 3.1) and in-the-wild personalization (Sec. 3.2).

3.1. Universal Prior Model for Clothed Humans

Geometry/Texture Representations. We represent the geometry and texture of clothed humans as a set of 3D Gaussians defined in a canonical space shared across identities. Each Gaussian is denoted as:

$$\mathbf{g} = \{\mathbf{x}, \mathbf{q}, \mathbf{s}, o, \mathbf{c}\}. \tag{1}$$

The parameters include a position $\mathbf{x} \in \mathbb{R}^3$, a unit quaternion $\mathbf{q} \in \mathbb{R}^4$, a scale factor $\mathbf{s} \in \mathbb{R}^3_+$ along three orthogonal axis, an opacity value $o \in \mathbb{R}_+$ and a color $\mathbf{c} \in \mathbb{R}^3_+$. The associated covariance matrix can be calculated as $\mathbf{\Sigma} = \mathbf{R}\,\mathrm{diag}(\mathbf{s})\,\mathrm{diag}(\mathbf{s})^\top \mathbf{R}^\top$, where the rotation matrix $\mathbf{R}$ can be easily converted from $\mathbf{q}$.

We deform the canonical 3D Gaussians to the posed space via forward linear blend skinning (LBS). Given the human pose $\boldsymbol{\theta}$ and shape $\boldsymbol{\beta}$ (consolidated as $\mathbf{\Theta}$), defined analogously to SMPL-X [57], we compute the forward transformation matrix $\mathbf{T}$ as follows:

$$\mathbf{T} = \sum_{i=1}^{n_b} w_c^i \mathbf{B}_i, \tag{2}$$

where n_b denotes the number of bones in the transformation and $\mathbf{B}_i$ are the bone transformation matrices for joints $i \in \{1, ..., n_b\}$, derived from the parameters $\mathbf{\Theta}$. $\mathbf{w}_c = \{w_c^1, ..., w_c^{n_b}\}$ represents the skinning weights which are queried from a diffused skinning map $\mathcal{S}_c$ based on the canonical SMPL-X vertices [47, 57]. The position $\mathbf{x}_c$ and covariance $\mathbf{\Sigma}_c$ of canonical Gaussians can then be transformed to the deformed space accordingly as follows:

$$\mathbf{x}_d = \mathbf{T}\mathbf{x}_c, \;\mathbf{\Sigma}_d = \mathbf{T}_{1:3}\mathbf{\Sigma}_c\mathbf{T}_{1:3}^\top. \tag{3}$$

Universal Identity Conditioning. We propose front and back maps as a universal parameterization for clothed humans. Such parameterization is obtained by orthogonally projecting the normalized canonical textured template into front and back views. Thus, we can obtain the front and back position maps $\mathcal{P}_c^{\mathrm{f}}$ and $\mathcal{P}_c^{\mathrm{b}}$, and texture maps $\mathcal{T}_c^{\mathrm{f}}$ and $\mathcal{T}_c^{\mathrm{b}}$ in canonical space. $\mathcal{P}_c^{\mathrm{f}}$ and $\mathcal{P}_c^{\mathrm{b}}$ are stored with the 3D positions of the canonical template. We abbreviate the concatenation of the canonical front and back position and texture maps as $\mathcal{P}_c$ and $\mathcal{T}_c$. For the sake of simplicity and clarity, we also adopt such abbreviation for other front and back maps. We use $\mathcal{T}_c$ as the identity conditioning data, enabling the feasibility of training a universal prior model $\mathcal{F}$ in a cross-identity setting. To accompany this, we define the output of our universal prior model $\mathcal{F}$ as a Gaussian map (front/back) $\mathcal{G}(\mathbf{\Theta})$, in which every pixel within the projected template mask $\mathcal{M}_c$ represents a 3D Gaussian with its attributes $\mathbf{g}$.

To obtain our conditioning data, we reconstruct the canonical template with vertices $\mathbb{V}$ via SDF-based volume rendering and LBS-based inverse mapping [15, 86] from multi-view images. The reconstruction is based on a single automatically selected keyframe in which the human pose $\boldsymbol{\theta}$ has the highest similarity to our pre-defined canonical human pose $\boldsymbol{\theta}_{\text{cano}}$. The texture is then obtained by unwrapping the RGB colors from the multi-view images onto the canonical template mesh.

To facilitate better learning of pose-dependent surface deformations and appearance changes across large-scale human identities, irrespective of bone lengths or human scales, we normalize the canonical templates based on an average human skeleton scale. Specifically, we set the shape parameters $\boldsymbol{\beta} = \mathbf{0}$ and pose parameters $\boldsymbol{\theta} = \boldsymbol{\theta}_{\text{cano}}$ for all subjects and compute the normalization transformation matrix $\mathbf{T}_{\text{norm}}$ similar to Eq. (2). We then apply the normalization to the canonical templates as follows:

$$\mathbf{v}_{\text{norm}} = \mathbf{T}_{\text{norm}}\mathbf{v}, \forall\, \mathbf{v} \in \mathbb{V}. \tag{4}$$

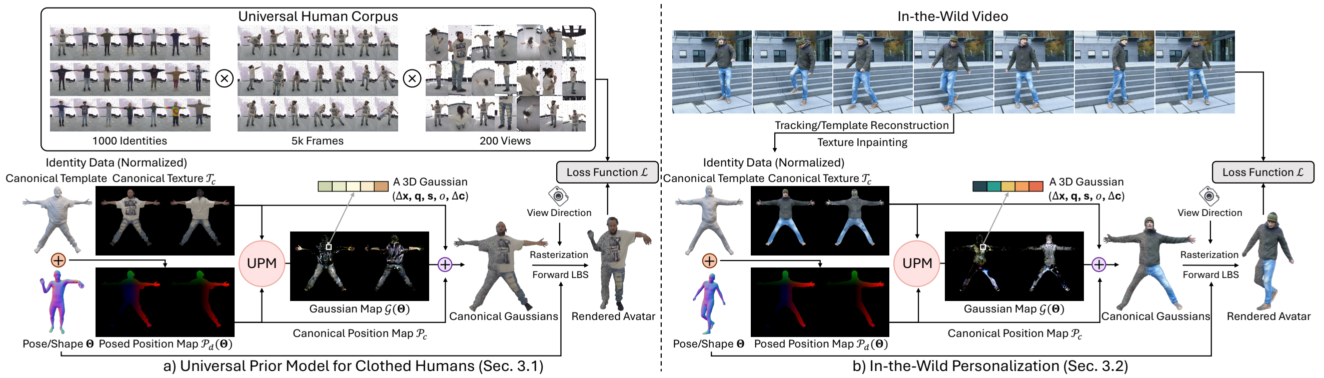

Figure 2. **Method Overview.** a) We employ a large corpus of multi-view dynamic clothed human performances to train a cross-identity universal prior model (UPM). During training, UPM is conditioned on the normalized identity-specific texture map $\mathcal{T}_c$ and takes the posed position map $\mathcal{P}_d(\mathbf{\Theta})$ as input to predict Gaussian attributes. We extract the canonical 3D Gaussians and synthesize human rendering for training pose/shape parameters $\mathbf{\Theta}$ by applying forward LBS and rasterization. We minimize the loss $\mathcal{L}$ over the entire universal human corpus. b) Given a monocular in-the-wild video of an unseen identity, we track the human pose/shape parameters $\mathbf{\Theta}$ and reconstruct the canonical textured template. We further deploy a diffusion-based model tailored for canonical texture inpainting to complete the canonical texture map. We then fine-tune our pre-trained UPM on the monocular observations via inverse rendering to recover person-specific details.

Given the normalized canonical textured template, we generate spatially aligned identity conditioning data for all training subjects by orthogonally projecting the templates into front and back views, yielding the canonical template masks $\mathcal{M}_c$, position maps $\mathcal{P}_c$, and texture maps $\mathcal{T}_c$.

Universal Prior Model Training. Inspired by prior work [5, 41], we employ a U-Net as the universal model backbone [64]. In particular, our universal model $\mathcal{F}$ takes the person-specific canonical texture maps $\mathcal{T}_c$ as identity conditioning input. To model pose-dependent effects such as dynamically changing wrinkles on clothes, we further concatenate the posed position maps $\mathcal{P}_d(\mathbf{\Theta})$ as additional input for each training iteration to predict pose-dependent Gaussian maps $\mathcal{G}(\mathbf{\Theta})$. Similar to deforming the canonical 3D Gaussians to the posed space (*cf*. Eq. (3)), we obtain the posed position maps $\mathcal{P}_d(\mathbf{\Theta})$ by applying forward LBS to the 3D positions stored in each valid pixel of the canonical position maps $\mathcal{P}_c$, excluding the global orientation and translation which do not affect the human dynamic details. Thus, our universal model works as follows:

$$\mathcal{F}(\mathcal{T}_c, \mathcal{P}_d(\mathbf{\Theta})) \rightarrow \mathcal{G}(\mathbf{\Theta}). \tag{5}$$

We extract the Gaussian attributes $\mathbf{g}$ for all valid pixels within $\mathcal{M}_c$ from the pose-dependent Gaussian maps $\mathcal{G}(\mathbf{\Theta})$. In practice, we opt to predict the position and color offsets $\Delta\mathbf{x}(\mathbf{\Theta}), \Delta\mathbf{c}(\mathbf{\Theta})$ relative to the canonical maps instead of absolute position/color maps. This encourages our model to focus on learning finer-grained details within a given model capacity. Thus, we attain canonical 3D Gaussians with positions $\mathbf{x} = \hat{\mathbf{x}} + \Delta\mathbf{x}(\mathbf{\Theta})$ and colors $\mathbf{c} = \hat{\mathbf{c}} + \Delta\mathbf{c}(\mathbf{\Theta})$, where $\hat{\mathbf{x}}$ and $\hat{\mathbf{c}}$ are queried from $\mathcal{P}_c$ and $\mathcal{T}_c$.

After applying Eq. (3), we obtain all 3D Gaussians in the posed space. Following splatting-based rasterization [34], the pixel color $\mathbf{C}$ is obtained by the α-blending of N overlapping Gaussians that are depth-ordered in posed space:

$$\mathbf{C} = \sum_{i=1}^{N} \mathbf{c}^i \alpha^i \prod_{j=1}^{i-1} \left(1 - \alpha^j\right), \tag{6}$$

where the transparency α is evaluated using the 2D covariance projected from $\mathbf{\Sigma}$ and multiplied with the per-Gaussian opacity o. Please refer to [34] for more details.

Training Objectives. The training objectives for our universal prior model include L1 and perceptual losses [89] between the rendered and ground-truth images, along with an offset regularization loss:

$$\mathcal{L} = \mathcal{L}_1 + \lambda_{\text{lpips}}\mathcal{L}_{\text{lpips}} + \lambda_{\text{offset}}\mathcal{L}_{\text{offset}}, \tag{7}$$

where $\lambda_{\{\cdot\}}$ denote loss weights, and the regularization loss $\mathcal{L}_{\text{offset}}$ penalizes excessively large per-Gaussian offsets $\Delta\mathbf{x}(\mathbf{\Theta})$. We minimize the loss function $\mathcal{L}$ across all multi-view clothed human performance capture data.

3.2. In-the-Wild Personalization

Preprocessing. Given the monocular in-the-wild video, we first track the human shape and poses $\mathbf{\Theta}$ using an off-the-shelf SMPL-X estimator [73]. We then employ Sapiens [35] to predict the 2D keypoints which are used to further refine the initial pose/shape estimates by minimizing the 2D keypoint projection error. These 2D keypoints are served as point prompts for SAM [33, 36] to acquire the human masks. Given the pose/shape estimation and segmentation masks, we employ a method akin to that used for studio data to reconstruct the 3D canonical textured template. Similarly, we follow the skeleton-based normalization strategy to attain the spatially aligned identity conditioning data. These steps serve to mitigate the domain gap in conditioning data between in-the-wild video sequences and the multi-view training data of our universal prior model. More details can be found in the Supp. Mat.

Diffusion-based Texture Inpainting. Unlike the multi-view capture data, human performance in in-the-wild videos exhibits varying levels of visibility, rather than consistent full-body visibility. Hence, we develop a latent diffusion model to inpaint the textures of canonical templates.

Instead of inpainting the textured template in 3D, we formulate this problem as a 2D inpainting task applied to the unwrapped canonical texture maps. We begin by generating various visibility masks through rasterization of the pre-acquired studio canonical templates using randomly positioned sparse cameras. We then fine-tune a pre-trained latent diffusion model with a DiT-based [58] architecture similar to that of [12] using the generated visibility masks and canonical texture maps derived from the studio data. Specifically, we train our diffusion model to denoise the canonical texture maps given partially masked inputs and their corresponding visibility masks. During inference, we obtain the visibility mask by rasterizing the template using the monocular input views and use it to generate a complete canonical texture map. Note that our diffusion model jointly operates on the front and back texture maps and can inpaint missing regions with plausible textures by utilizing context from both sides. We refer to the Supp. Mat for more details.

Fine-Tuning. To create 3D avatars with identity-preserving details while maintaining the generalization power to novel human poses, we further fine-tune our universal prior model on the monocular video observations via inverse rendering with a small number of iterations. Due to the potentially inaccurate pose initialization, we fine-tune all network parameters, and optimize the estimated human shape and per-frame pose parameters $\boldsymbol{\Theta}$ jointly. The loss function employed during the fine-tuning stage mirrors that utilized in the universal model training stage (*cf*. Eq. (7)).

4. Experiments

We first introduce the training corpus and the test datasets. Then we compare Vid2Avatar-Pro with state-of-the-art approaches in two tasks: interpolation and extrapolation synthesis. Ablation studies are then conducted to show the effectiveness of our core components and design choices. Finally, we demonstrate more animation results of avatars created from in-the-wild monocular videos qualitatively.

Table 1. **Quantitative interpolation synthesis comparisons on NeuMan dataset.** Our method consistently outperforms all baselines on all evaluation metrics (*cf*. Fig. 3).

Method	PSNR ↑	SSIM ↑	LPIPS ↓
HumanNeRF [78]	27.06	0.967	1.92
InstantAvatar [28]	28.47	0.972	2.77
NeuMan [29]	25.48	0.966	2.87
Vid2Avatar [15]	26.87	0.969	2.41
3DGS-Avatar [61]	29.75	0.975	1.75
GaussianAvatar [21]	29.94	0.980	1.24
ExAvatar [52]	31.39	0.981	1.64
Ours	32.71	0.983	1.19

4.1. Training Dataset

We use a multi-view system similar to [2] to capture dynamic human performance, where we obtain calibrated and synchronized multi-view images at a resolution of 4096×2668 pixels through the use of 200 cameras. Participants are instructed to perform casual human motions for an average of approximately 5,000 frames per person. In total, 1,000 participants are recorded for model training.

4.2. Test Datasets

NeuMan Dataset [29]: This dataset comprises a collection of videos captured by a mobile phone, featuring a single person performing a walking motion. Following previous works [21, 52], we use *bike*, *citron*, *jogging*, and *seattle* sequences for comparisons that exhibit most body regions and contain minimally blurry images. Since the official training and testing splits have interwoven frames, we regard the evaluation on NeuMan dataset as an interpolation view synthesis task.

MonoPerfCap Dataset [84]: This dataset contains in-the-wild videos of people with different garment types and various daily actions. Compared to NeuMan dataset [29], MonoPerfCap dataset includes more clothing deformations and pose diversity. We use the first 80% of each captured video for training and the remaining 20% frames for testing. Thus, we define the evaluation on MonoPerfCap dataset as a more challenging extrapolation view synthesis task.

Evaluation Protocol: We report PSNR, SSIM [76], and LPIPS (×100) [89] for all synthesis comparisons. In all evaluation tables, the top three techniques are highlighted in red, orange, and yellow, respectively.

4.3. Interpolation Synthesis Comparisons

We conduct interpolation synthesis comparisons on NeuMan dataset [29]. For methods that jointly model the human and the background, we only compare the foreground (human) rendering quality. The quantitative results are partially

Figure 3. **Qualitative interpolation synthesis comparisons on NeuMan dataset.** Baseline methods tend to render with artifacts (*e.g.*, corrupted faces and feet) and less details (*e.g.*, the clothing wrinkles and the clothes zipper). In contrast, our method generates clean and realistic human renderings while recovering more appearance details (*e.g.*, facial features).

Table 2. **Quantitative extrapolation synthesis comparisons on MonoPerfCap dataset.** Our method consistently outperforms all baselines on all evaluation metrics (*cf*. Fig. 4).

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Vid2Avatar [15]	28.49	0.976	2.46
ExAvatar [52]	30.29	0.979	2.19
Ours w/o Fine-Tuning	29.24	0.977	2.29
Ours w/o Avg. Scale	31.03	0.979	1.51
Ours	31.97	0.981	1.37

sourced from [21, 52]. Overall, our method outperforms all baseline methods in the interpolation setting by a substantial margin on all metrics (*cf*. Tab. 1). The difference is more visible in the qualitative comparison shown in Fig. 3, where baseline methods cannot produce sharp details such as the wrinkles on the clothing and zippers, and tend to generate artifacts on the face or feet. In contrast, Vid2Avatar-Pro recovers more surface details (*e.g.*, clothing wrinkles and facial features). We attribute this to our expressive universal prior model that is trained on high-quality data.

4.4. Extrapolation Synthesis Comparisons

We further conduct a more challenging experiment: extrapolation synthesis comparisons on MonoPerfCap dataset [84]. Despite a more challenging setting, Tab. 2 indicates that our method still outperforms baselines across all metrics. This disparity becomes more evident in qualitative comparisons presented in Fig. 4. ExAvatar [52] struggles to recover the accurate facial features, and both ExAvatar [52] and Vid2Avatar [15] produce blurry renderings (*cf*. the shoes in the first row of Fig. 4). Thanks to our pre-trained universal prior model, Vid2Avatar-Pro can generalize better to out-of-distribution poses with plausible pose-dependent surface/appearance deformations (*cf*. the hem in the second row of Fig. 4). Our method also consistently preserves superior photorealism compared to the baselines.

Table 3. **Number of training IDs.** Rendering quality consistently increases when the prior model is trained on more identities.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Ours #ID 4	31.28	0.979	1.53
Ours #ID 16	31.28	0.980	1.51
Ours #ID 128	31.34	0.980	1.45
Ours #ID 1000 (Full)	31.97	0.981	1.37

4.5. Ablation Studies

Training Data. A key factor that affects the performance of our universal prior model is the number of training identities. We quantitatively analyze the effects by training the universal prior model with different numbers of training subjects, *i.e.* $n_p \in \{4, 16, 128, 1000\}$. The training subjects are chosen at random. We use MonoPerfCap dataset to evaluate the extrapolation synthesis performance. The quantitative results are reported in Tab. 3. We observe that the final rendering quality consistently improves as the amount of training data/identities increases.

Conditioning Data Normalization. To investigate the importance of skeleton-based conditioning data normalization, we compare our full model to a version without normalizing using an average scale (w/o Avg. Scale). We demonstrate in Tab. 2 that skeleton-based normalization helps to improve the generalization ability of our model to unseen identities and challenging poses (*cf*. Fig. 5).

Figure 4. **Qualitative extrapolation synthesis comparisons on MonoPerfCap dataset.** In this more challenging setting, our method produces more identity-preserving human renderings with finer-grained appearance details, *e.g.*, the facial features and shoes. Powered by our universal prior model, Vid2Avatar-Pro generates more plausible pose-dependent surface deformations (*cf*. the hem of the T-shirt).

Figure 5. **Importance of skeleton-based normalization.** Without skeleton-based conditioning data normalization, the created avatar tends to produce artifacts when driven with out-of-distribution poses (*e.g.*, the holes close to the armpit).

Figure 6. **Importance of inpainting.** Our diffusion-based inpainting module can effectively complete the textures that are missing from the monocular observations.

Diffusion-based Texture Inpainting. We introduce a diffusion-based canonical inpainting module to complete the missing textures that are not visible in the monocular in-the-wild videos. As illustrated in Fig. 6, our inpainting module effectively eliminates artifacts caused by missing observations and inpaint the 3D avatar with plausible textures. More examples can be found in the Supp. Mat.

Fine-Tuning. Due to the inherent domain gap between in-the-wild videos and studio data, fine-tuning the prior model based on monocular observations is crucial to recovering person-specific details. To demonstrate the effectiveness, we conduct an ablation experiment without fine-tuning on MonoPerfCap dataset (w/o Fine-Tuning). Results in Tab. 2 indicate that without fine-tuning, the final renderings are blurrier and lack detailed appearance features, as shown in Fig. 8 (*e.g.*, the T-shirt pattern and the belt).

4.6. Qualitative Results

In Fig. 7, we showcase the animation results of avatars created from in-the-wild monocular videos. Vid2Avatar-Pro can generalize to different identities and garment styles, and produce highly realistic renderings for novel human poses and arbitrary view points. More qualitative results animated by various poses are available in the Supp. Mat.

Figure 7. **Visual animation results of avatars created from monocular in-the-wild videos.** The created 3D avatars can be animated using novel human poses and demonstrate highly detailed appearance from arbitrary view points.

Figure 8. **Importance of fine-tuning.** Without fine-tuning, the human avatar cannot preserve the fine-grained human appearance details such as the T-shirt pattern and the belt.

5. Conclusion

In this work, we present Vid2Avatar-Pro to create photorealistic and animatable 3D avatars from monocular in-the-wild videos via universal prior. We first propose a universal parameterization for clothed humans that enables cross-identity training schema. We then build a photorealistic and animatable universal prior model for clothed humans learned from a thousand of high-quality dynamic performance capture data. Finally, we construct a robust pipeline that supports personalization of the universal prior model to monocular in-the-wild videos and create personalized photorealistic human avatars that can be faithfully animated to novel human motions and rendered from novel view points.

Limitations and Future Works: The current training dataset for our universal prior model lacks diverse facial expressions and dynamic capture data of human subjects dressed in loose garments. Therefore, our method currently does not support animatable faces or loose clothing. Additionally, our method assumes standard lighting conditions and may not perform optimally in environments with extreme lighting variations. We refer to Supp. Mat for a more detailed discussion of limitations and societal impact.

References

[1] Thiemo Alldieck, Mihai Zanfir, and Cristian Sminchisescu. Photorealistic monocular 3d reconstruction of humans wearing clothing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[2] Timur Bagautdinov, Chenglei Wu, Tomas Simon, Fabián Prada, Takaaki Shiratori, Shih-En Wei, Weipeng Xu, Yaser Sheikh, and Jason Saragih. Driving-signal aware full-body avatars. *ACM Transactions on Graphics (TOG)*, 40(4):1–17, 2021. 1, 2, 5

[3] Marcel C. Buehler, Gengyan Li, Erroll Wood, Leonhard Helminger, Xu Chen, Tanmay Shah, Daoye Wang, Stephan Garbin, Sergio Orts-Escolano, Otmar Hilliges, Dmitry Lagun, Jérémy Riviere, Paulo Gotardo, Thabo Beeler, Abhimitra Meka, and Kripasindhu Sarkar. Cafca: High-quality novel view synthesis of expressive faces from casual few-shot captures. In *ACM SIGGRAPH Asia 2024 Conference Paper*, 2024. 3

[4] Marcel C Bühler, Kripasindhu Sarkar, Tanmay Shah, Gengyan Li, Daoye Wang, Leonhard Helminger, Sergio Orts-Escolano, Dmitry Lagun, Otmar Hilliges, Thabo Beeler, et al. Preface: A data-driven volumetric prior for few-shot ultra high-resolution face synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3402–3413, 2023.

[5] Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, Yaser Sheikh, and Jason Saragih. Authentic volumetric avatars from a phone scan. *ACM Trans. Graph.*, 41(4), 2022. 2, 3, 4

[6] Aggelina Chatziagapi, Grigorios G. Chrysos, and Dimitris Samaras. Migs: Multi-identity gaussian splatting via tensor decomposition. In *ECCV*, 2024. 3

[7] Jianchuan Chen, Wentao Yi, Liqian Ma, Xu Jia, and Huchuan Lu. Gm-nerf: Learning generalizable model-based neural radiance fields from multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20648–20658, 2023. 3

[8] Mingfei Chen, Jianfeng Zhang, Xiangyu Xu, Lijuan Liu, Yujun Cai, Jiashi Feng, and Shuicheng Yan. Geometry-guided progressive nerf for generalizable and efficient neural human rendering. In *ECCV*, 2022. 3

[9] Yushuo Chen, Zerong Zheng, Zhe Li, Chao Xu, and Yebin Liu. Meshavatar: Learning high-quality triangular human avatars from multi-view videos. In *ECCV*, 2024. 2

[10] Zhaoxi Chen, Gyeongsik Moon, Kaiwen Guo, Chen Cao, Stanislav Pidhorskyi, Tomas Simon, Rohan Joshi, Yuan Dong, Yichen Xu, Bernardo Pires, He Wen, Lucas Evans, Bo Peng, Julia Buffalini, Autumn Trimble, Kevyn McPhail, Melissa Schoeller, Shoou-I Yu, Javier Romero, Michael Zollhöfer, Yaser Sheikh, Ziwei Liu, and Shunsuke Saito. URhand: Universal relightable hands. In *CVPR*, 2024. 2, 3

[11] Enric Corona, Tomas Hodan, Minh Vo, Francesc Moreno-Noguer, Chris Sweeney, Richard Newcombe, and Lingni Ma. Lisa: Learning implicit shape and appearance of hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20533–20543, 2022. 3

[12] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 2, 5

[13] Qiao Feng, Yebin Liu, Yu-Kun Lai, Jingyu Yang, and Kun Li. Fof: Learning fourier occupancy field for monocular real-time human reconstruction, 2023. 2

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

[15] Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, and Otmar Hilliges. Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 3, 5, 6

[16] Chen Guo, Tianjian Jiang, Manuel Kaufmann, Chengwei Zheng, Julien Valentin, Jie Song, and Otmar Hilliges. Reloo: Reconstructing humans dressed in loose garments from monocular video in the wild. In *European conference on computer vision (ECCV)*, 2024. 2

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

[18] Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Real-time deep dynamic characters. *ACM Transactions on Graphics*, 40 (4), 2021. 2

[19] Tong He, Yuanlu Xu, Shunsuke Saito, Stefano Soatto, and Tony Tung. Arch++: Animation-ready clothed human reconstruction revisited. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11046–11056, 2021. 2, 3

[20] Hsuan-I Ho, Jie Song, and Otmar Hilliges. Sith: Single-view textured human reconstruction with image-conditioned diffusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3

[21] Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 5, 6

[22] Shoukang Hu, Tao Hu, and Ziwei Liu. Gauhuman: Articulated gaussian splatting from monocular human videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20418–20431, 2024. 1, 2

[23] Tao Hu, Tao Yu, Zerong Zheng, He Zhang, Yebin Liu, and Matthias Zwicker. Hvtr: Hybrid volumetric-textural rendering for human avatars. In *2022 International Conference on 3D Vision (3DV)*, 2022. 2

[24] Yangyi Huang, Hongwei Yi, Yuliang Xiu, Tingting Liao, Jiaxiang Tang, Deng Cai, and Justus Thies. TeCH: Text-guided Reconstruction of Lifelike Clothed Humans. In *International Conference on 3D Vision (3DV)*, 2024. 2

[25] 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, 3

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

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

[28] 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 (CVPR)*, 2023. 5

[29] 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. 1, 5

[30] Zeren Jiang, Chen Guo, Manuel Kaufmann, Tianjian Jiang, Julien Valentin, Otmar Hilliges, and Jie Song. Multiply: Reconstruction of multiple people from monocular video in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[31] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[32] HyunJun Jung, Nikolas Brasch, Jifei Song, Eduardo Perez-Pellitero, Yiren Zhou, Zhihao Li, Nassir Navab, and Benjamin Busam. Deformable 3d gaussian splatting for animatable human avatars, 2023. 2

[33] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. In *NeurIPS*, 2023. 4

[34] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. 2, 4

[35] Rawal Khirodkar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *Computer Vision – ECCV 2024*, pages 206–228, Cham, 2025. Springer Nature Switzerland. 4

[36] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 4

[37] Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. HUGS: Human gaussian splatting. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[38] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. In *Advances in Neural Information Processing Systems*, 2021. 3

[39] Youngjoong Kwon, Baole Fang, Yixing Lu, Haoye Dong, Cheng Zhang, Francisco Vicente Carrasco, Albert Mosella-Montoro, Jianjin Xu, Shingo Takagi, Daeil Kim, Aayush Prakash, and Fernando De la Torre. Generalizable human gaussians for sparse view synthesis. In *European Conference on Computer Vision*, 2024. 3

[40] Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. Gart: Gaussian articulated template models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19876–19887, 2024. 2

[41] Junxuan Li, Chen Cao, Gabriel Schwartz, Rawal Khirodkar, Christian Richardt, Tomas Simon, Yaser Sheikh, and Shunsuke Saito. Uravatar: Universal relightable gaussian codec avatars. In *ACM SIGGRAPH 2024 Conference Papers*, 2024. 2, 3, 4

[42] Mingwei Li, Jiachen Tao, Zongxin Yang, and Yi Yang. Human101: Training 100+fps human gaussians in 100s from 1 view, 2023. 2

[43] Mengtian Li, Shengxiang Yao, Zhifeng Xie, and Keyu Chen. Gaussianbody: Clothed human reconstruction via 3d gaussian splatting, 2024. 2

[44] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhofer, Jurgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[45] Zhe Li, Zerong Zheng, Yuxiao Liu, Boyao Zhou, and Yebin Liu. Posevocab: Learning joint-structured pose embeddings for human avatar modeling. *ACM SIGGRAPH Conference Proceedings*, 2023. 2

[46] Zhe Li, Zerong Zheng, Lizhen Wang, and Yebin Liu. Animatable gaussians: Learning pose-dependent gaussian maps for high-fidelity human avatar modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[47] Siyou Lin, Hongwen Zhang, Zerong Zheng, Ruizhi Shao, and Yebin Liu. Learning implicit templates for point-based clothed human modeling. In *ECCV*, 2022. 3

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

[49] Xinqi Liu, Chenming Wu, Jialun Liu, Xing Liu, Jinbo Wu, Chen Zhao, Haocheng Feng, Errui Ding, and Jingdong Wang. Gva: Reconstructing vivid 3d gaussian avatars from monocular videos, 2024. 2

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

[51] 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 *European conference on computer vision*, pages 405–421. Springer, 2020. 2

[52] Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3D gaussian avatar. In *ECCV*, 2024. 1, 2, 5, 6

[53] Gyeongsik Moon, Weipeng Xu, Rohan Joshi, Chenglei Wu, and Takaaki Shiratori. Authentic hand avatar from a phone scan via universal hand model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2029–2038, 2024. 3

[54] Arthur Moreau, Jifei Song, Helisa Dhamo, Richard Shaw, Yiren Zhou, and Eduardo Pérez-Pellitero. Human gaussian splatting: Real-time rendering of animatable avatars. In *CVPR*, 2024. 2

[55] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *International Conference on Computer Vision*, 2021. 2

[56] Haokai Pang, Heming Zhu, Adam Kortylewski, Christian Theobalt, and Marc Habermann. Ash: Animatable gaussian splats for efficient and photoreal human rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1165–1175, 2024. 2

[57] 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 *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 1, 2, 3

[58] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 5

[59] 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. 2

[60] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021. 2

[61] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2, 5

[62] 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

[63] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *IEEE International Conference on Computer Vision Workshops*, 2021. 2

[64] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. 2, 4

[65] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2304–2314, 2019. 2, 3

[66] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020. 2, 3

[67] Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. Relightable gaussian codec avatars. In *CVPR*, 2024. 2

[68] Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. SplattingAvatar: Realistic Real-Time Human Avatars with Mesh-Embedded Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[69] Kaiyue Shen, Chen Guo, Manuel Kaufmann, Juan Zarate, Julien Valentin, Jie Song, and Otmar Hilliges. X-avatar: Expressive human avatars. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

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

[71] Shih-Yang Su, Timur Bagautdinov, and Helge Rhodin. Danbo: Disentangled articulated neural body representations via graph neural networks. In *European Conference on Computer Vision*, 2022. 2

[72] Guoxing Sun, Rishabh Dabral, Pascal Fua, Christian Theobalt, and Marc Habermann. Metacap: Meta-learning priors from multi-view imagery for sparse-view human performance capture and rendering. In *ECCV*, 2024. 3

[73] Qingping Sun, Yanjun Wang, Ailing Zeng, Wanqi Yin, Chen Wei, Wenjia Wang, Haiyi Mei, Chi-Sing Leung, Ziwei Liu, Lei Yang, and Zhongang Cai. Aios: All-in-one-stage expressive human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1834–1843, 2024. 4

[74] David Svitov, Pietro Morerio, Lourdes Agapito, and Alessio Del Bue. Haha: Highly articulated gaussian human avatars with textured mesh prior, 2024. 2

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

[76] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers*, 2003. 5

[77] Jing Wen, Xiaoming Zhao, Zhongzheng Ren, Alex Schwing, and Shenlong Wang. GoMAvatar: Efficient Animatable Human Modeling from Monocular Video Using Gaussians-on-Mesh. In *CVPR*, 2024. 2

[78] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220, 2022. 1, 2, 5

[79] Donglai Xiang, Fabian Prada, Timur Bagautdinov, Weipeng Xu, Yuan Dong, He Wen, Jessica Hodgins, and Chenglei Wu. Modeling clothing as a separate layer for an animatable human avatar. *ACM Trans. Graph.*, 40(6), 2021. 2

[80] Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. ICON: Implicit Clothed humans Obtained from Normals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13296–13306, 2022. 2, 3

[81] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal integration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[82] Yuliang Xiu, Yufei Ye, Zhen Liu, Dimitrios Tzionas, and Michael J Black. Puzzleavatar: Assembling 3d avatars from personal albums. *ACM Transactions on Graphics (TOG)*, 2024. 2

[83] Tianhan Xu, Yasuhiro Fujita, and Eiichi Matsumoto. Surface-aligned neural radiance fields for controllable 3d human synthesis. In *CVPR*, 2022. 2

[84] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. Monoperfcap: Human performance capture from monocular video. *SIGGRAPH*, 37(2):27:1–27:15, 2018. 2, 5, 6

[85] Lixin Xue, Chen Guo, Chengwei Zheng, Fangjinhua Wang, Tianjian Jiang, Hsuan-I Ho, Manuel Kaufmann, Jie Song, and Hilliges Otmar. HSR: holistic 3d human-scene reconstruction from monocular videos. In *European Conference on Computer Vision (ECCV)*, 2024. 2

[86] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Advances in Neural Information Processing Systems*, 2021. 2, 3

[87] Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Zarate, Jie Song, and Otmar Hilliges. Hi4d: 4d instance segmentation of close human interaction. In *Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[88] Zhengming Yu, Wei Cheng, Xian Liu, Wayne Wu, and Kwan-Yee Lin. Monohuman: Animatable human neural field from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16943–16953, 2023. 1, 2

[89] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 4, 5

[90] Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. Humannerf: Efficiently generated human radiance field from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7743–7753, 2022. 3

[91] Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[92] Yang Zheng, Qingqing Zhao, Guandao Yang, Wang Yifan, Donglai Xiang, Florian Dubost, Dmitry Lagun, Thabo Beeler, Federico Tombari, Leonidas Guibas, and Gordon Wetzstein. Physavatar: Learning the physics of dressed 3d avatars from visual observations. In *European Conference on Computer Vision (ECCV)*, 2024. 2

[93] Zerong Zheng, Tao Yu, Yebin Liu, and Qionghai Dai. Pamir: Parametric model-conditioned implicit representation for image-based human reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[94] Zerong Zheng, Xiaochen Zhao, Hongwen Zhang, Boning Liu, and Yebin Liu. Avatarrex: Real-time expressive full-body avatars. *ACM Transactions on Graphics (TOG)*, 42(4), 2023. 2

[95] Heming Zhu, Fangneng Zhan, Christian Theobalt, and Marc Habermann. Trihuman: A real-time and controllable tri-plane representation for detailed human geometry and appearance synthesis. *ACM Trans. Graph.*, 2024. 2

[96] Wojciech Zielonka, Timur Bagautdinov, Shunsuke Saito, Michael Zollhöfer, Justus Thies, and Javier Romero. Drivable 3d gaussian avatars, 2023. 2