DOI: 10.1111/cgf.70192
Eurographics Symposium on Geometry Processing 2025
M. Attene and S. Sellán
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 44 (2025), Number 5

Representing Animatable Avatar via Factorized Neural Fields

Chunjin Song[1], Zhijie Wu[1], Bastian Wandt[3], Leonid Sigal[1,2], Helge Rhodin[1,4†]

[1]University of British Columbia [2]Vector Institute for AI [3]Linköping University [4]Bielefeld University

Abstract

For reconstructing high-fidelity human 3D models from monocular videos, it is crucial to maintain consistent large-scale body shapes along with finely matched subtle wrinkles. This paper explores how per-frame rendering results can be factorized into a pose-independent component and a corresponding pose-dependent counterpart to facilitate frame consistency at multiple scales. Pose adaptive texture features are further improved by restricting the frequency bands of these two components. Pose-independent outputs are expected to be low-frequency, while high-frequency information is linked to pose-dependent factors. We implement this with a dual-branch network. The first branch takes coordinates in the canonical space as input, while the second one additionally considers features outputted by the first branch and pose information of each frame. A final network integrates the information predicted by both branches and utilizes volume rendering to generate photo-realistic 3D human images. Through experiments, we demonstrate that our method consistently surpasses all state-of-the-art methods in preserving high-frequency details and ensuring consistent body contours. Our code is accessible at `https://github.com/ChunjinSong/facavatar`.

CCS Concepts

• ***Computing methodologies*** → *Reconstruction; Shape inference;*

1. Introduction

Neural body models [HLX*23, KLF*23, PZK*24] now yield personalized, almost photorealistic 3D human avatars even from a single 2D video [WCS*22, JYS*22, YCL*23, GJC*23]. Recently, they have garnered significant attention from both academia and industry due to their broad range of downstream applications in virtual reality, gaming, and e-commerce. The most widespread models learn a Neural Radiance Field (NeRF) [SYZR21, SBR22, WCS*22] or a set of 3D Gaussians [QWM*24, KCG*24, MSD*24, LWP*24] by conditioning the underlying neural network on the input pose. In detail, existing models impose constraints by learning the neural field relative to the skeleton and restricting pose-dependent changes, e.g., to be local through a GNN [NSLH21, SYZR21, SBR23, SWR24]. However, they risk overfitting, as a high-capacity model may learn to render training views without reconstructing a temporally consistent 3D shape. In turn, they are typically parameterized to smooth over high-frequency details leading to artifacts in shape and texture, even with explicitly chosen frequency-based representation [SWR24].

The use of a canonical reference frame [JYS*22, LTV*22, WSGT22] in which the shape and appearance of the character are defined statically, e.g. in a T-pose, is an alternative direction to the models described above that predict in the observation space per-frame. A 3D model can be learned from videos by deforming the single canonical model to the pose visible in every frame with explicit surface meshes [XCZ*18], Gaussian representations [RRR*15, RRC*16], or NeRF based [SYZR21, SBR22, WCS*22] methods. They either explicitly learn a mapping [SYZR21, WCS*22, YCL*23] or implicitly use root finding [LTV*22, WSGT22] to enable a warp field from observation to canonical space. However, learning and constraining this deformation also has generalization difficulties due to the inferior frame consistency.

Fig. 1 gives an overview of our method. Taking a monocular video as input, our goal is to learn an accurate human representation which preserves consistent body shapes along with finely matched details. To achieve this, after mapping positions in observation spaces to the canonical space through skeletal deformation [WCS*22, GJC*23], previous methods attempt to either directly compute geometry and appearance outputs with canonical positions and per-frame poses [GJC*23, SBR23] (sketch in Fig. 2 (a)), or explicitly estimate pose-dependent deformation in canonical coordinate spaces [WCS*22] (sketch in Fig. 2 (b)). While these, in principle, are good ideas, they cannot maintain frame consistency and adaptive fine-grained details simultaneously due to lacking constraints on the highly non-linear outputs from neural networks. In contrast, inspired by works on signal decompositions [SWZ*19, WSC*22] and private-shared component analysis in other contexts [SEUD10, BTS*16, WSZ*20] which explicitly extract common features to enhance generalization, our model decomposes the network outputs into pose-independent and pose-dependent components. In Fig. 2 (c), we employ a neural network to explicitly extract pose-independent outputs with canoni-

† Corresponding author

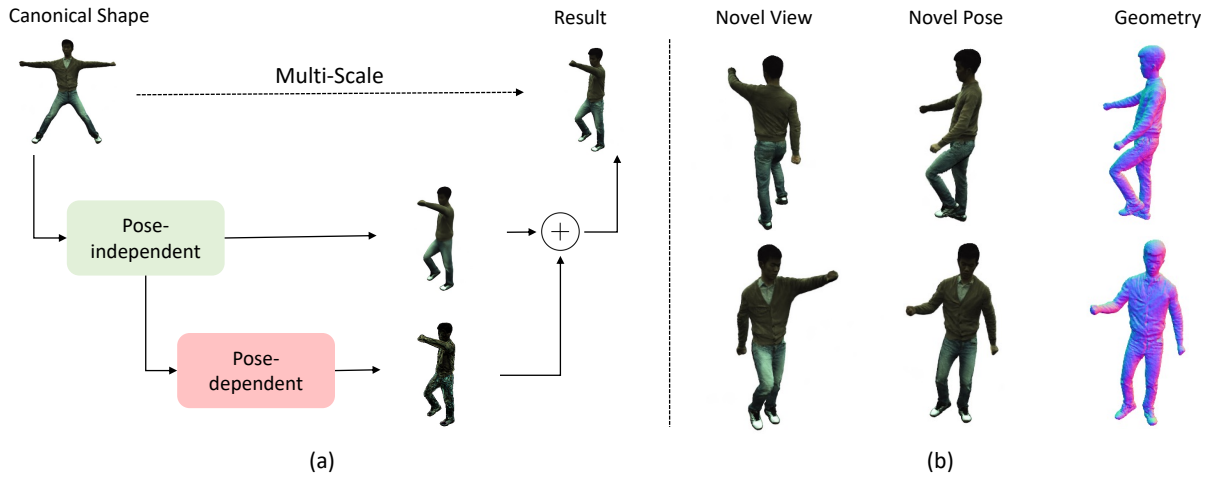

Figure 1: ***Motivation illustration.*** *(a) In canonical space, we separate the per-frame rendering output into a pose-independent component and its pose-dependent equivalent. These two components are modeled with distinct frequency bands, thus yielding smooth base outputs and corresponding high-frequency residuals (see Fig. 4 for details). The residual image here is amplified for better visualization. (b) Our frequency-aware factorization improves the state-of-the-art methods in novel view synthesis, novel pose rendering and human shape reconstruction.*

cal coordinates. These operations are shared across all frames and enable pose-independent information, which represents the common characteristics of the subject throughout the video, to directly enhance frame consistency and advance network generalization.

As shown in Fig. 4, increasing frequencies in the pose-independent output facilitates more geometric details. However, it prevents the full model from synthesizing sharper pose-dependent wrinkles as the high frequency leaves the shared representations vulnerable to contamination by specific pose-dependent patterns [SEUD10, BTS*16], making the model harder to generalize. We thus enforce the per-frame high-frequency details to only be modeled by pose-dependent output factors. Fig. 1 (a) illustrates the effect of this frequency assumption. The pose-independent and pose-dependent components characterize the smooth shared shape contours and adaptive high-frequency residuals respectively. The final result is computed by mixing these two components.

Moreover, we follow [YGKL21, GJC*23] to enable detailed geometry modeling by applying the Signed Distance Function (SDF) for NeRF-based volume rendering. We also design an objective for the pose-independent branch to encourage as much information as possible to be encoded in the shared outputs for improved generalization. Our contributions can be summarized as:

- Introducing a novel neural network to generate high-fidelity human representations via the frequency-aware factorized field.
- Designing a simple objective for the pose-independent branch that maximizes the shared information to improve generality.
- Demonstrating significant improvements on the state-of-the-art methods in novel view synthesis, novel pose rendering and shape reconstruction simultaneously.

2. Related Work

In the two following sections, we review related methods in neural field modeling [XTS*22] and delve into the latest and most relevant approaches in neural avatar modeling.

Neural Fields. Due to the impressive performance, neural fields [PFS*19, WWL*19, PSSS19, MON*19, GCV*19, GCS*20, TLY*21, MESK22] are extensively studied to enhance their generalization [YYTK21], compactness [WJMY23, RLN*23], level of detail [TLY*21, MESK22, TET*22], camera self-calibration [LMTL21, YCFB*21], and resource efficiency [MESK22, SSC22]. A notable advancement in neural fields is Neural Radiance Fields (NeRF) [MST*20], developed for rendering images from arbitrary camera views in static scenes. Subsequent efforts have extended NeRF to dynamic scenes [GSKH21, LSZ*22, LNSW21, DZY*21, PSB*21, PSH*21, TTG*21, YLSL21] and apply the multi-band NeRF methods [LVVPW22, SLFB22] to explicitly extract multi-scale frequency components for static scenes. However, all of these methods do not take human body priors (e.g. skeleton) into account and do not address significant time-dependent non-rigid deformations commonly encountered in learning human avatar representations [SYZR21, SBR22, LTV*22, OMT*21, PZX*21, XAS21]. Some other works attempt to apply the Signed Distance Function (SDF) for NeRF-based models to extract accurate 3D shapes [YGKL21, WLL*21, WSW22] but not for neural avatar modeling as well.

Neural Fields for Avatar Modeling. In textured avatar modeling, the parametric SMPL body model serves as a common foundation [ZHY*22, ZZZ*23]. Conversely, approaches such as A-NeRF [SYZR21] and NARF [NSLH21] lack a surface prior, directly transforming input query points into relative coordinates of skeletal joints. TAVA [LTV*22] and ARAH [WSGT22] use root finding to enable a warp field from observation to canonical space. Inspired by the disentanglement learning [HLBK18, LTH*18, SWZ*19, WSZ*20, WSC*22, SZP*23, WCT*24], DANBO [SBR22] and NPC [SBR23] model the per-part feature space using a graph neural network for better scalability. Moreover, some methods [LHR*21, PZX*21, PDW*21, DFG*22] aim to improve results with an image-to-image translation network and a per-frame latent code while others [GPX*23, JCSH23] try to model dynamic humans based on learnable grid structures [MESK22] and enable more effi-

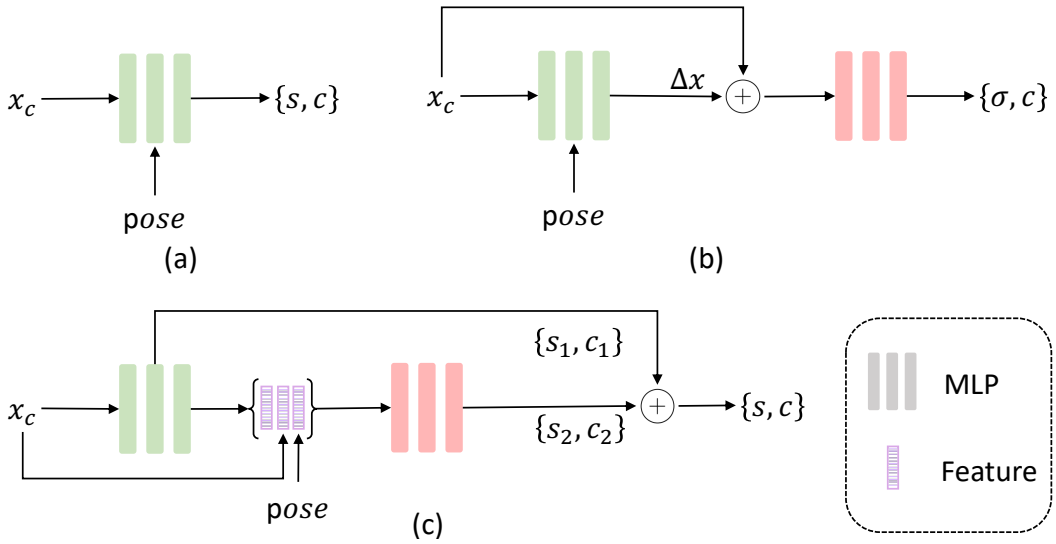

Figure 2: ***Conceptual differences.*** *Taken a position x_c in canonical space and conditioned on a pose, Vid2Avatar [GJC*23] directly regresses the SDF and appearance values $\{s, c\}$ with a uniform frequency band and thus models pose-independent information implicitly (a). In (b), HumanNeRF [WCS*22] and MonoHuman [YCL*23] perform decomposition for* ***input coordinates in the canonical space*** *to enable non-rigid deformation. In comparison, we associate the pose-independent* ***network outputs*** *$\{s_1, c_1\}$ with low frequencies and pose-dependent counterparts $\{s_2, c_2\}$ with high-frequencies to preserve multi-scale signals (c). Here x_c is computed by a skeletal deformation [WCS*22, GJC*23] as described in Sec. 3.1. MLP networks colored in red and green correspond to high and low frequencies respectively.*

cient training and inference. For the human modeling from monocular videos, [WCS*22, YCL*23, GJC*23] learn a canonical space for all frames to improve frame consistencies and testing generalization. Recently, some works [QWM*24, KCG*24, MSD*24, LWP*24, HZZ*24] apply the Gaussian Splatting framework [KKLD23] for better inference speed, but they lack geometric detail. Differing from all these methods, we explicitly separate a rendering image into a pose-independent factor and pose-dependent counterpart to improve frame consistency and synthesize adaptive details.

Fig. 2 further illustrates the conceptual differences between our method and the two most closely related baselines that incorporate deformations into the canonical space. Both Vid2Avatar [GJC*23] and our method apply SDF to possess constant shape contours and improve frame consistencies. However, Vid2Avatar directly maps the input positions at the canonical space and poses to geometry and appearance with uniform frequencies and does not model the pose-invariant information for better generalization. Thus it either introduces unwanted artifacts or blurs the desired high-frequency textures (e.g. 2^{nd} row in Fig. 6). Both HumanNeRF [WCS*22] and LS-Avatar [SWS*25] decompose the deformation into a rigid pose-independent part and its non-rigid pose-dependent counterpart in the canonical coordinate space. While they can synthesize fine-grained details, outputs from the highly non-linear MLP twist the body shape with fuzzy boundaries as shown in the 1^{st} row of Fig. 6. In contrast, our method has superior frame consistency in body outlines and better generalization to precise textures due to explicit pose-independent modeling and accurate frequency associations of network outputs.

Moreover, both PM-Avatar [SWR24] and our method conceptually utilize frequency modeling to improve human avatar modeling. However, the main idea of PM-Avatar is to modulate query points' frequency transformation based on part-level pose contexts while we improve baselines with a frequency-aware factorization strategy for network outputs. In detail, we differ as follows: 1) We explicitly decompose outputs into pose-independent and pose-dependent components, whereas PM-Avatar implicitly models pose-independent components; 2) We associate these components with low and high frequencies respectively while PM-Avatar uses uniform frequency bands for modulation; 3) We input whole skeletons for simplicity while PM-Avatar focuses on part-level pose modeling; 4) While PM-Avatar only models pose-dependent features, we additionally leverage a canonical space for better generalization; 5) Instead of density-based rendering, we adopt an SDF representation to enable surface-based rendering and improve shape reconstruction. These differences enable superior robustness in monocular settings.

3. Method

We aim to reconstruct a 3D animatable avatar by leveraging a collection of N images. Fig. 3 provides a method overview with three main components. First, we compute a query point x_o in the per-frame observation space as $x_o = \mathbf{O} + t * \mathbf{d}$, where $\mathbf{O}$ is the position of the camera center, $\mathbf{d}$ is the ray direction and t indicates the distance along the ray. Then we estimate the body pose for an input frame, which is represented as the sequence of joint angles $[\theta_k]_{k=1}^{N_B}$. These joint angles are used to perform the skeletal deformation for x_o and obtain the coordinate x_c in canonical space. Next the computed position x_c is passed to the two-branch network to output the pose-independent SDF and color values and the pose-dependent counterparts. Finally, we merge the pose-independent and pose-dependent outputs, which

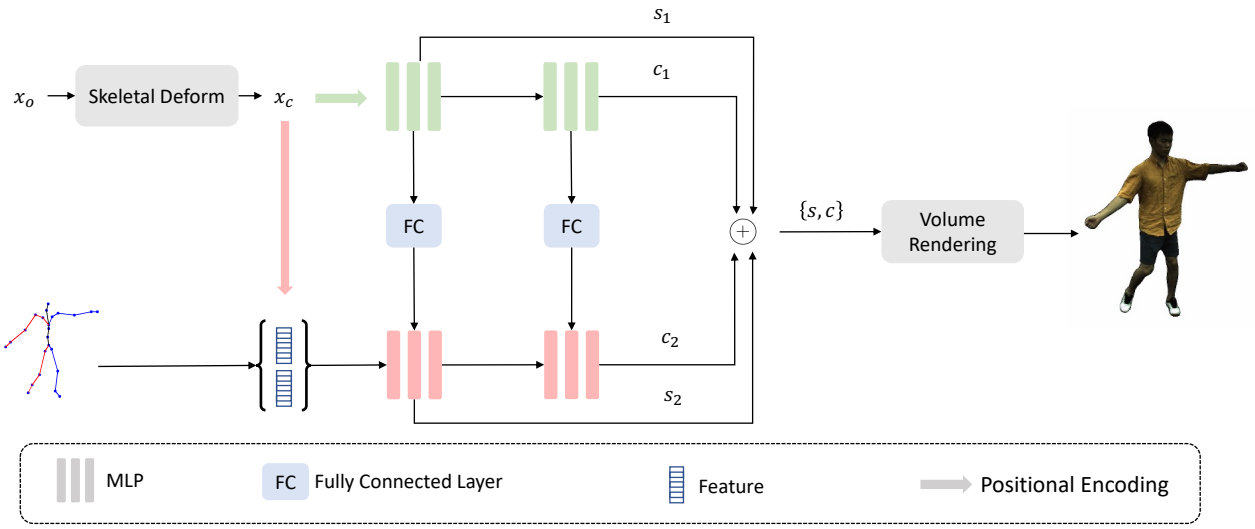

Figure 3: ***Architecture overview***. *We compute the canonical coordinate x_c of the query point x_o in observation space by performing the skeletal deformation. Then x_c is fed into two branches with the low-frequency (green) and high-frequency (red) positional encoding for pose-independent ($\{s_1, c_1\}$) and pose-dependent ($\{s_2, c_2\}$) outputs respectively. We input their combinations $\{s, c\}$ to volume rendering to generate images under different view directions and human poses.*

are then mapped to the corresponding density and radiance at that location as in the SDF-based NeRF framework.

3.1. Skeletal Deformation

Modeling 3D avatars in a canonical space is crucial to form a temporally consistent representation. We follow Vid2Avatar [GJC*23] to perform the skeletal transformation from x_o in observation to x_c in canonical space. Given the N_B joint angles $\theta = [\mathbf{B}_1, \ldots, \mathbf{B}_{N_B}]$ of the given body pose, the inverse of linear-blend skinning computes

$$x_c = \left(\sum_{i=1}^{N_B} \omega_i \mathbf{B}_i \right)^{-1} x_o, \tag{1}$$

where $\{\omega_i\}$ denotes the skinning weight of the i-th bone and is based on the point-to-point distances to the nearest SMPL vertices in observation space; see [GJC*23] for details. Compared to the learnable skinning weights [WCS*22], this SMPL-based weights can significantly reduce GPU memory consumption. This procedure also stabilizes the network training with faster convergence speed.

3.2. Factorized Neural Fields

Our main contribution is to factorize the rendering results of animatable avatars into pose-independent and pose-dependent parts and associate these components with low and high frequencies, respectively. To achieve this, we design a novel factorized two-branch network as shown in Fig. 3. Taken the canonical position x_c as input, we first feed x_c into a low-frequency positional encoding [MST*20]

$$\begin{aligned} \gamma^1(x_c) = (x_c, \sin(2^0 \pi x_c), \cos(2^0 \pi x_c), \ldots, \\ \sin(2^{L_{ind}-1} \pi x_c), \cos(2^{L_{ind}-1} \pi x_c)), \end{aligned} \tag{2}$$

where L_{ind} indicates the highest mapping frequency in $\gamma^1(x_c)$. Then the processed feature $\gamma^1(x_c)$ is fed to the upper branch θ^{ind}, which consists of two MLPs with ReLU activations, to estimate the pose-independent geometry $\{s_1\}$ and appearance $\{c_1\}$ components. To leverage the relationships between pose-independent and pose-dependent information and synthesize adaptive details, the upper branch also learns to output a feature vector for the SDF network and color network on the bottom branch respectively

$$s_1, f_{\text{sdf}}^1 = \theta_{\text{sdf}}^{ind}\left(\gamma^1(x_c)\right); \quad c_1, f_c^1 = \theta_c^{ind}\left(f_{\text{sdf}}^1, \overrightarrow{n}_1\right). \tag{3}$$

Here $\overrightarrow{n}_1$ indicates the normalized gradient of s_1 computed at x_c. $\theta_{\text{sdf}}^{ind}$ and θ_c^{ind} denote the SDF network and color network of the upper branch, respectively.

As mentioned above, we observe that the pose-dependent patterns, such as dynamically changing wrinkles on clothes, should process higher frequencies than its pose-independent analogue. Thus, we feed x_c into another high-frequency positional encoding as:

$$\begin{aligned} \gamma^2(x_c) = (x_c, \sin(2^0 \pi x_c), \cos(2^0 \pi x_c), \ldots, \\ \sin(2^{L_d-1} \pi x_c), \cos(2^{L_d-1} \pi x_c)). \end{aligned} \tag{4}$$

Similar to L_{ind} in $\gamma^1(x_c)$, L_d here stands for the highest transformation frequency of $\gamma^2(x_c)$. We then give the concatenation of $\gamma^2(x_c)$, f_{sdf}^1 and the input pose θ to the bottom branch θ^d to output pose-dependent geometry $\{s_2\}$ and appearance $\{c_2\}$ components as

$$s_2, f_{\text{sdf}}^2 = \theta_{\text{sdf}}^d\left([\gamma^2(x_c), f_{\text{sdf}}^1, \theta]\right), \quad c_2 = \theta_c^d\left([f_{\text{sdf}}^2, f_c^1, \overrightarrow{n}, \theta]\right). \tag{5}$$

To combine the low-frequency pose-independent and high-frequency pose-dependent outputs, we compute the final SDF and color outputs for further processing as

$$s = s_1 + s_2, \quad c = c_1 + c_2. \tag{6}$$

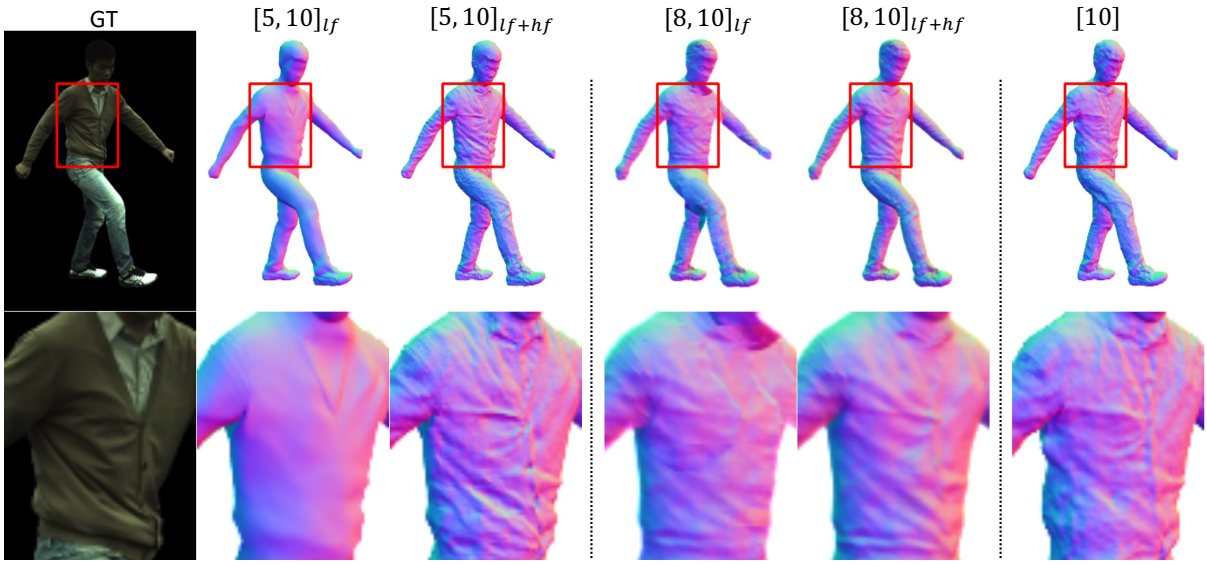

Figure 4: ***Frequency constraints.*** *To validate our frequency assumption, we train a set of two-branch models with different L_{ind} and L_d. For simplicity, we denote a model with $L_{ind} = x$ and $L_d = y$ as $[x,y]$. Adhering to our network design, the pose-independent branch outputs the low-frequency base normal map as $[5,10]_{lf}$ while our full model estimates an output with all frequencies as $[5,10]_{lf+hf}$. Increasing frequency in the pose-independent output, denoted as $[8,10]$, can yield more grainy geometric patterns in $[8,10]_{lf}$ but stops the full model from generating sharp pose-dependent wrinkles in $[8,10]_{lf+hf}$. Simply training the pose-dependent branch ($[10]$ with $L_d = 10$) fails to synthesize desired multi-scale patterns. See Sec. 3.2 and Sec. B in appendix for more discussions.*

Note that we apply the gradient vector $\overrightarrow{n}$ of the learned signed distance function s to the color network to facilitate the disentanglement of geometry and appearance [YGKL21]. Although s represents the final geometry signals as SDF values, both s_1 and s_2 do not necessarily encode meaningful geometric information. We set $L_d = 10$ and halve $L_{ind} = 5$ for simplicity across all experiments.

Discussions on frequency bands. We enforce the high-frequency information to only be encoded in pose-dependent output for improving the synthesis of pose adaptive results. A few high-frequency patterns, such as the facial structures and shoe textures in Fig. 4, only deform lightly, and could be encoded into a pose-independent output with high frequency. However, we only use one single positional encoding setting for all parts and such part-dependent properties are not known as prior and may change from part to part. We trade off how much information each part contributes to the shared features. As shown in Fig. 4, simply increasing frequency helps create more realistic facial patterns and introduce more geometric details in the highlighted torso. But it also goes against the generalization to novel poses and leads to blunt pose-dependent wrinkles as in $[8,10]_{lf+hf}$. Similar to Vid2Avatar, the one-branch network with uniform frequency band fails to reproduce multi-scale geometries. Moreover, the distinct frequency assignments help penalize redundant information between the pose-independent and pose-dependent branches. See Tab. I in the appendix for additional quantitative results.

3.3. SDF-based volume rendering

We utilize a SDF-based representation opposed to classical NeRF or Gaussian Splatting as it provides the most faithful geometry. With the output SDF and color signals $\{s,c\}$, we first compute the density via a learnable transformation [YGKL21], namely

$$\sigma = \frac{1}{\beta} \cdot \Psi_\beta(-s), \tag{7}$$

where β is a learnable parameter and Ψ is the Cumulative Distribution Function (CDF) of the Laplace distribution with zero mean and β scale.

Following the existing neural radiance rendering pipelines for human avatars [SYZR21, SBR22, WSGT22, SWR24], we output the image of a subject for a ray r as in the original NeRF:

$$\hat{C}(r) = \sum_{i=1}^{n} \mathcal{T}_i (1 - \exp(-\sigma_i \delta_i))\, c_i, \mathcal{T}_i = \exp\left(- \sum_{j=1}^{i-1} \sigma_j \delta_j \right). \tag{8}$$

Here, $\hat{C}$ and δ_i indicate the synthesized image and the distance between adjacent samples along a given ray respectively. Finally, we compute the L_1 loss $\|\cdot\|_1$ for training as

$$\mathcal{L}_{\text{rec}} = \sum_{\mathbf{r} \in \mathfrak{R}} \left\| \hat{C}(r) - C_{gt}(r) \right\|_1, \tag{9}$$

where $\mathfrak{R}$ is the whole ray set and C_{gt} is the ground truth.

Following [YGKL21, GJC*23], we apply the Eikonal loss to satisfy the Eikonal equation such that the learned $s = s_1 + s_2$ can approximate a signed distance function as

$$\mathcal{L}_{\text{eik}} = \mathbb{E}_{x_c} \left(\left\| \frac{\mathrm{d}s}{\mathrm{d}x_c} \right\| - 1 \right)^2. \tag{10}$$

We observe that it is crucial to maximize the amount of information in the pose-independent (upper) branch to improve generality.

To avoid that the upper branch degrades and the pose-dependent branch takes most information, we render the human image by volume rendering with the pose-independent SDF and color components $\{s_1, c_1\}$ and compute the L_1 loss for common data as

$$\mathcal{L}_{\text{com}} = \sum_{\mathbf{r}\in\mathfrak{R}} \left\| \bar{C}(r) - C_{gt}(r) \right\|_1, \tag{11}$$

where $\bar{C}$ is the synthesized image created by $\{s_1, c_1\}$.

Besides the aforementioned L_1 loss, a perceptual loss, LPIPS [ZIE*18], is employed to provide robustness to slight misalignments and shading variation and to improve detail in the reconstruction as $\mathcal{L}_{\text{LPIPS}}$.

Thus, given a video sequence of a human, we aim to optimize the following combined loss function:

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_{\text{eik}}\mathcal{L}_{\text{eik}} + \lambda_{\text{com}}\mathcal{L}_{\text{com}} + \lambda_{\text{LPIPS}}\mathcal{L}_{\text{LPIPS}}. \tag{12}$$

Here λ_{eik}, λ_{com} and λ_{LPIPS} are weights for Eikonal loss, common loss and LPIPS loss, respectively.

4. Results

In this section, we compare our approach with the state-of-the-art methods for rendering and 3D shape reconstruction, comparing to canonical and observation space based methods: HumanNeRF [WCS*22], MonoHuman [YCL*23], NPC [SBR23], Vid2Avatar [GJC*23], PM-Avatar [SWR24], 3DGS-Avatar [QWM*24] and GoMAvatar [WZR*24]. We also conduct ablation studies to analyze and discuss the effects of factorized avatar representation, common loss function and the dependencies between pose-independent and pose-dependent branches. Source code will be released upon publication.

4.1. Experimental Settings

We assess the effectiveness of our approach using well-established benchmarks for body modeling. Following HumanNeRF and MonoHuman, we conduct evaluations across the eight sequences of the ZJU-Mocap dataset [PZX*21]. To measure the geometry reconstruction, we also adopt SynWild examples [GJC*23] as testing protocol. Additionally, we utilize two publicly available YouTube sequences as an in-the-wild dataset to gauge performance on everyday monocular videos. We follow Vid2Avatar to process all sequences and obtain approximate camera and body pose with off-the-shelf estimators. We employ the SAM model [KMR*23] to generate foreground maps of all images for precise segmentation.

To ensure comparison fairness, we follow former experimental settings, including dataset splits and metrics [WCS*22, YCL*23, SBR23]. Our evaluation covers standard image metrics like pixel-wise Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Metric (SSIM) [WBSS04] to assess image quality. We also utilize perceptual metrics such as Learned Perceptual Image Patch Similarity (LPIPS) [ZIE*18], Kernel Inception Distance (KID) [BSAG18], and Fréchet Inception Distance (FID) [HRU*17] to evaluate structural accuracy and textured details of generated images. All metrics are computed across entire generated images. And both LPIPS and KID metrics are multiplied by 1000 for better comparisons. Similar to ARAH [WSGT22], we additionally report Chamfer Distance (CD) and Normal Consistency (NC) to estimate the quality of reconstructed shapes.

Table 1: ***Novel-view and novel-pose synthesis results on the ZJU-Mocap test set.*** *For both novel view synthesis and novel pose rendering, our method notably improves baselines in LPIPS and KID which are reported to be more informative than PSNR and SSIM in the monocular setting [QWM*24].*

	Novel view				Novel pose			
	PSNR↑	SSIM↑	LPIPS↓	KID↓	PSNR↑	SSIM↑	LPIPS↓	KID↓
HumanNeRF	29.66	0.967	36.78	14.23	29.57	0.966	34.17	12.32
MonoHuman	30.18	0.967	31.45	13.18	29.90	0.967	32.21	12.61
NPC	30.01	0.967	37.18	53.24	29.61	0.967	36.52	49.79
PM-Avatar	30.27	0.969	38.38	39.64	29.87	0.969	39.26	40.16
Vid2Avatar	29.76	0.969	35.61	27.65	29.53	0.969	35.69	31.51
3DGS-Avatar	30.09	0.968	31.30	15.33	29.77	0.968	30.69	13.24
GoMAvatar	**30.29**	0.968	32.40	12.80	**30.20**	0.968	32.03	13.81
Ours	30.11	**0.970**	**29.64**	**11.72**	29.98	**0.970**	**28.60**	**11.09**

4.2. Novel View Synthesis

We utilize ZJU-Mocap sequences as a multi-view dataset to evaluate the generalization capability under different camera views. Specifically, we use images captured by the first provided camera ("camera 1") for training and the remaining images for evaluation.

We visualize the results in the upper row of Fig. 6 for NeRF based methods and those in the left part of Fig. 7 for 3DGS based methods respectively. Comparing to baselines, our method shows superior capabilities in recovering fine-grained details (e.g. the vertical patterns). Additionally, our method better preserves the body shape, such as cloth contours. We attribute this to the explicit separation in output space which mitigates grainy artifacts and reserves consistent large-scale outlines. We compare quantitatively in Tab. 1 to further support our previous findings.

Table 2: *We report L_2 Chamfer Distance (CD) and Normal Consistency (NC) over the ZJU-Mocap and SynWild sequences for geometry reconstruction evaluation. Our model shows better overall perceptual quality and shape reconstruction from monocular videos. *The imperfect pseudo-ground-truth smooths details, benefiting Vid2Avatar for the NC metric. Despite the clear visual improvements, the inaccurate pose estimation prevents fully convincing results on SynWild examples. See Sec. 4.4 and appendix for details.*

	ZJU-Mocap		SynWild		Average	
	CD↓	NC↑	CD↓	NC↑	CD↓	NC↑
HumanNeRF	0.242	0.649	0.182	0.667	0.212	0.658
MonoHuman	0.318	0.636	0.229	0.642	0.274	0.639
NPC	0.079	0.795	0.217	0.640	0.148	0.718
PM-Avatar	0.054	0.770	0.212	0.666	0.133	0.718
Vid2Avatar	0.053	**0.878***	0.196	**0.741**	0.125	**0.808**
3DGS-Avatar	0.316	0.612	0.232	0.586	0.274	0.599
GoMAvatar	0.048	0.823	0.201	0.715	0.125	0.769
Ours	**0.047**	0.863	**0.180**	0.731	**0.113**	0.798

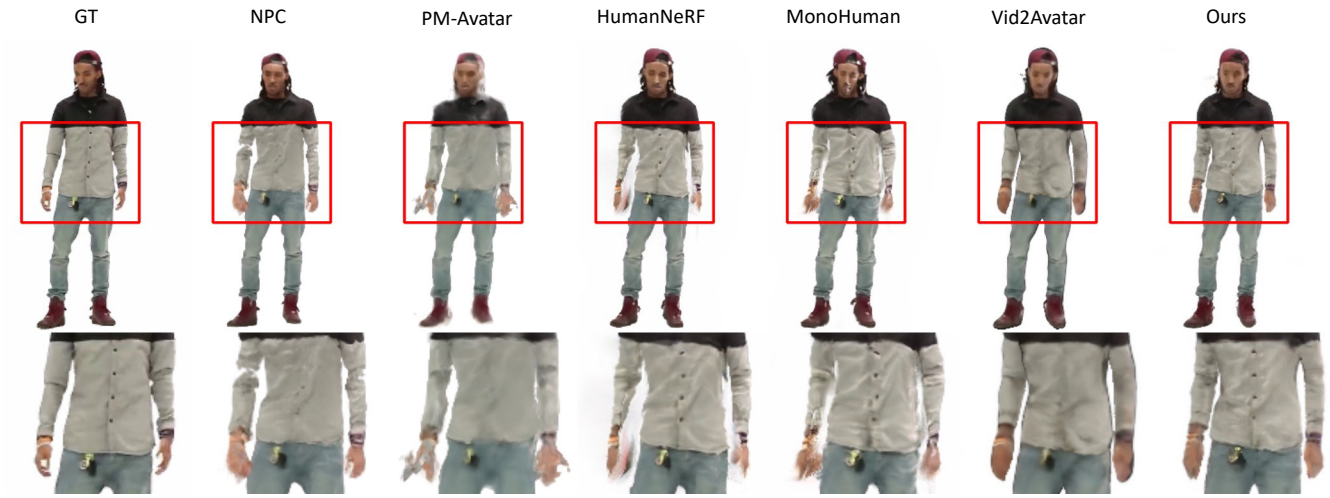

Figure 5: ***Novel pose rendering on Youtube sequences.*** *While baselines distort the marked arms with floating noise, our method yields more visually appealing body outlines. We also improve Vid2Avatar with more realistic textures like cloth buttons.*

4.3. Novel Pose Rendering

Rendering under novel poses is critical to many down-streamed applications like computer animations. To evaluate the generalization to unseen human poses, we train all models on the first part of a video and test on the remaining frames. During evaluation, only the 3D human pose is used as input.

Also on the ZJU-Mocap dataset, Fig. 6 and the right part of Fig. 7 present the visual comparisons for the NeRF based and 3DGS based methods respectively. Our method demonstrates superior results, with sharper body boundaries (e.g., arms) and more detailed textures (e.g., wrinkles). In contrast, baselines either introduce blurry patterns (e.g. Vid2Avatar) or distort the shape contours with noisy artifacts (e.g. MonoHuman). Fig. 5 additionally presents the results of Youtube sequences. While our method succeeds in generating reasonable multi-scale patterns, baselines severely distort the highlighted arms under challenging poses. Tab. 1 and Tab. 3 further quantitatively verify our strong generalization to unseen poses. Although we achieve slightly superior or comparable PSNR and SSIM metrics to baselines, here we mainly apply the perceptual metrics, LPIPS and FID, to evaluate outdoor sequences as they are more robust to the varying lighting conditions than pixel-based alternatives as suggested by NPC [SBR23], Note that none of the methods achieve perfect alignment with wrinkle locations due to their chaotic formation on unseen poses. Addressing this issue using physics falls outside the scope of this paper. Notably, generalizing to novel poses is more challenging than novel view synthesis and the case where we attain the largest relative improvements compared to baselines. This empirical evidence substantiates the effectiveness of our frequency-aware factorized avatar representation.

Generating realistic wrinkles in response to varying human motions is essential for neural avatar methods. To illustrate this, Fig. 8 shows the synthesis of dynamic cloth wrinkles across different frames. Specifically, the T-shirt wrinkles, highlighted by the red arrow, appear as diagonal stripes that gradually fade and shift direction as the subject moves. These smooth wrinkle dynamics demonstrate our method's adaptability to diverse input poses.

4.4. Geometry Comparisons

In Fig. 9, we analyze the 3D meshes reconstructed with our approach against reconstructions from the baselines. Our method better captures the smooth body surfaces and detailed geometry (e.g. the wrinkles). In contrast, the baselines predict more noisy blobs near the body surface whose structured patterns (e.g. facial expressions) cannot be faithfully recognized. While Vid2Avatar does provide a complete body outline, it tends to overly smooth out sharp textures and generate flat patterns. The 'ZJU-Mocap' column in Tab. 2 additionally complies with our empirical advantages in a quantitative manner. Note that we follow ARAH [WSGT22] to compute pseudo ground truth which by itself smooths surface details, more than our method does. Fig. 10 and the 'SynWild' column in Tab. 2 provide the visual demonstration and quantitative comparisons on the SynWild dataset correspondingly. Being consistent with better generalization of novel view synthesis and novel pose rendering, the improvements of geometry reconstruction suggest that more precise modeling of geometry is beneficial for the visual fidelity.

Table 3: ***Unseen pose animation on Youtube videos.*** *Our model shows better perceptual quality from in-the-wild videos.*

	Story		Invisible		Average	
	LPIPS↓	FID↓	LPIPS↓	FID↓	LPIPS↓	FID↓
HumanNeRF	31.35	63.28	33.72	72.29	32.54	67.79
MonoHuman	32.73	65.23	34.39	79.94	33.56	72.59
NPC	29.59	53.62	35.28	80.17	32.44	66.90
PM-Avatar	35.15	78.86	42.67	109.49	38.91	94.18
Vid2Avatar	36.85	187.24	40.52	198.51	38.69	192.88
Ours	**28.52**	**56.57**	**31.74**	**69.13**	**30.13**	**62.85**

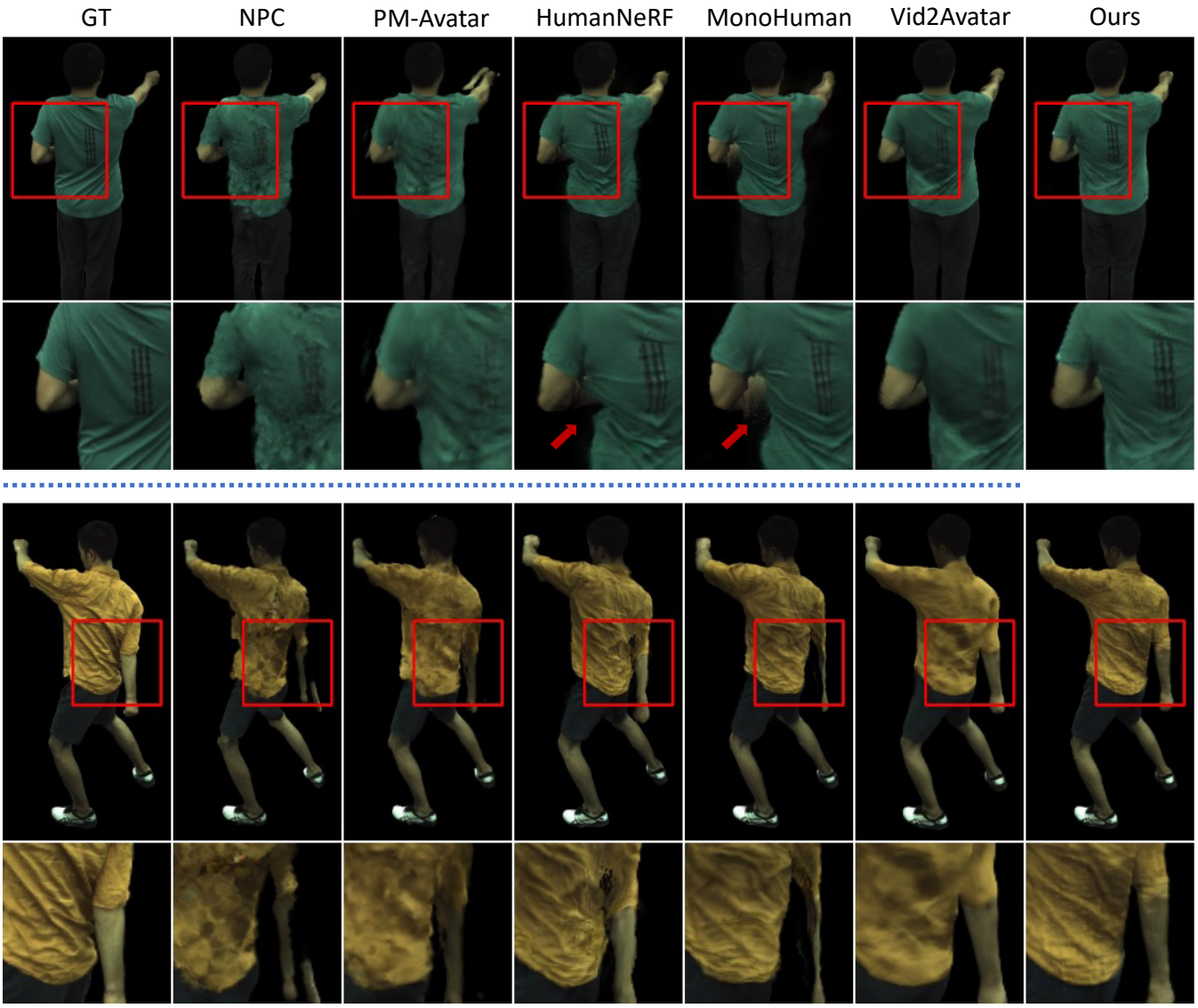

Figure 6: ***Visual comparisons with NeRF based methods on ZJU-Mocap.*** *Our method can render superior sharp contours and synthesize more adaptive textures under novel camera views (upper) and novel avatar poses (bottom).*

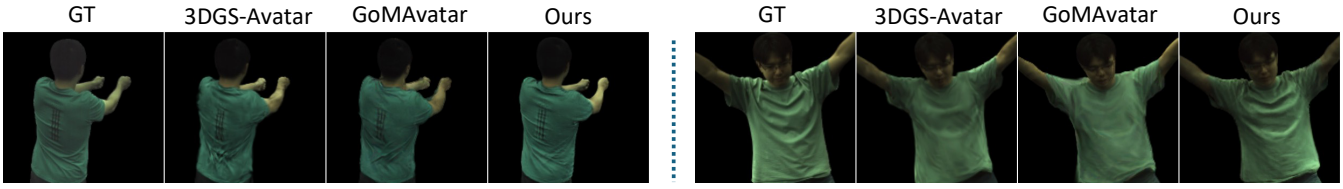

Figure 7: ***Comparisons with 3D Gaussian Splatting (3DGS) based methods on ZJU-Mocap.*** *Our method can more vividly reproduce structured patterns (e.g. the vertical pattern designs and cloth folds) without introducing distorted artifacts than baselines.*

4.5. Ablation studies

We conduct ablation studies with the following ablated models: **1.** Only preserve the bottom branch network with pose-dependent deformations as 'w/o $\{s_1, c_1\}$'; **2.** Only preserve the upper branch network with pose-independent deformations as 'w/o $\{s_2, c_2\}$'; **3.** We remove the common loss function as 'w/o $\mathcal{L}_{\text{com}}$'; **4.** For the bottom branch, we only input the target pose and x_c as 'w/o feat'; **5.** We feed the body pose to the pose-independent branch instead of the pose-dependent branch as Pose_{lf}. These experiments are performed

Figure 8: ***Dynamic wrinkle synthesis.*** *As highlighted by the red arrow, the generated wrinkles on the T-shirt, initially appeared as the diagonal stripes, gradually vanish and shift directions as human moves.*

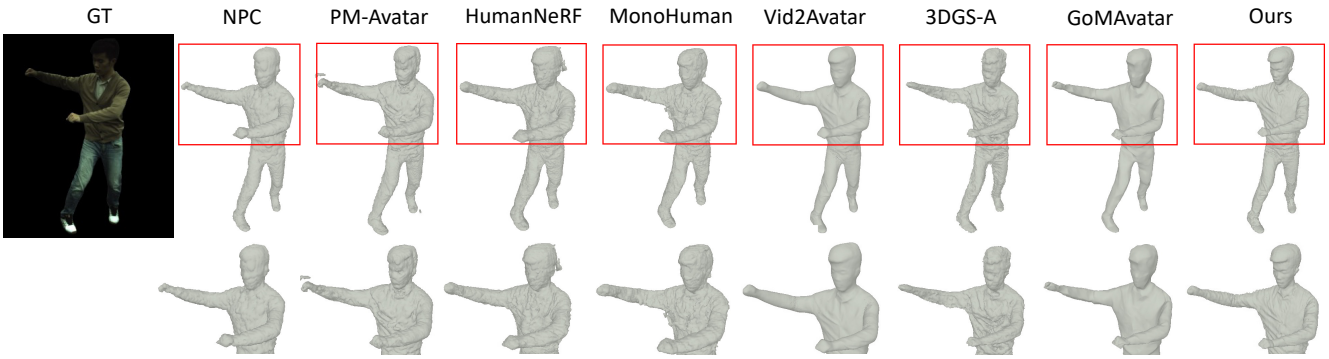

Figure 9: ***Geometry reconstructions on ZJU-Mocap.*** *Our factorized avatar representation achieves state-of-the-art performance in respect of seamless human surfaces and precise geometric details.*

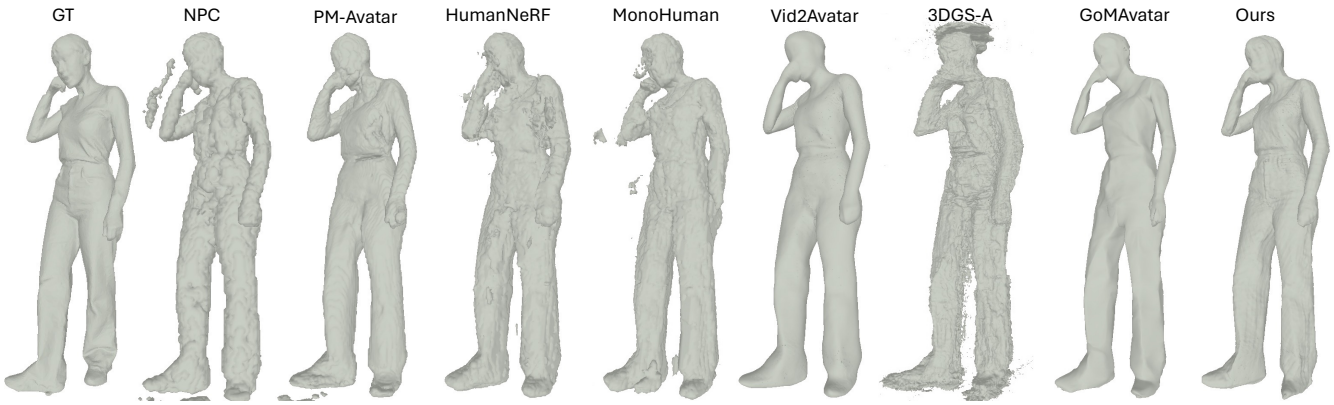

Figure 10: ***Geometry reconstructions on SynWild.*** *While most baselines synthesize bumpy surface outlines, Vid2Avatar and GoMAvatar overly smooth out the desired geometric variations, yielding generally flat surface patterns. In contrast, our method can produce better wrinkles and facial structures with reasonable large-scale appearances.*

to evaluate the effectiveness of the factorized fields, the common loss $\mathcal{L}_{\text{com}}$ and the dependency between two branches respectively. Pose$_{lf}$ is further trained to evaluate the importance of common information among frames. We perform comparisons in both novel view synthesis and novel pose rendering on the ZJU-Mocap S394 sequence. The quantitative results shown in Tab. 4 consistently highlight the importance of all network components. We additionally offer more ablation study results and discussions in the appendix.

5. Limitation and Discussions

Due to the dense MLP computation in the volume rendering framework, computation time remains a constraint for real-time applications. To address this issue, some works apply the grid-based implementation to constrain the representation computation in a local area [MESK22, CXG*22, WJMY23]. Our method requires individual training for each actor and cannot generalize to other humans without additional training. Training a generalizable human representation with foundation models is a promising direction. Since we do not explicitly consider pattern editing in our current

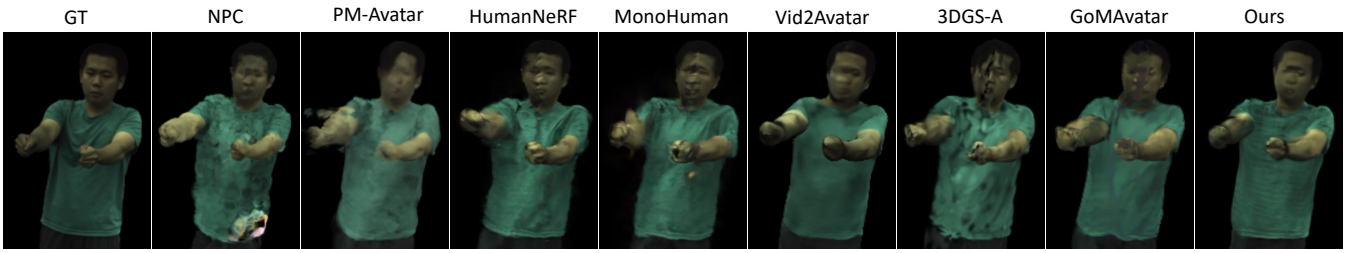

Figure 11: ***Failure cases on ZJU-Mocap.*** *How to generalize to challenging cases is still an open problem, where all methods fail to preserve the visually pleasing textures and body outlines under this pose.*

Figure 12: ***Failure cases on ActorsHQ.*** *While our method avoids fragmenting the subject's body or introducing floating artifacts in empty space, extending it to better model loose clothing is a promising area for future exploration.*

Table 4: ***Ablation study on ZJU-Mocap sequence.*** *Our full model outperforms all ablated baselines across all metrics.*

	Novel view			Novel pose		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
w/o $\{s_1, c_1\}$	30.13	0.965	31.00	29.36	0.962	31.01
w/o $\{s_2, c_2\}$	29.56	0.963	32.46	29.26	0.962	31.30
w/o $\mathcal{L}_{\text{com}}$	30.08	0.964	31.40	29.47	0.963	30.90
w/o feat	29.49	0.962	33.41	29.15	0.961	32.77
Pose_{lf}	30.11	0.965	29.83	29.28	0.963	30.53
Ours	**30.68**	**0.967**	**29.74**	**29.69**	**0.965**	**29.41**

framework design, how to enhance it with editing features is also our future work.

In Fig. 11, we demonstrate that, under extreme challenges, when the test pose is very different from the training poses, our method cannot fully adaptively reproduce the target textures but distorts the body contours. However, it is important to recognize that addressing these issues remains an open problem in neural avatar modeling.

Recently, there has been growing attention on modeling loose-fitting garments. Different from the tight cloth learning which assumes that both apparels and human skeleton follow the same motion trajectories, the movements of human body and loose cloths are weakly entangled. Thus we extend our empirical comparisons to ActorsHQ dataset [IRG*23] to verify the generalization of each method to this difficult case. Fig. 12 illustrates the rendering results. While our method is capable of synthesizing overall reasonable geometry contours, it blurs the cloth deformation which is an interesting question to address.

6. Conclusions

We introduce a novel two-branch framework to enhance the accuracy of avatar representation learning. Our primary contribution lies in a unique frequency-aware field factorization design, which enhances frame consistency and boosts the ability to produce adaptive details. In comparison to existing methods, our approach demonstrates empirical advantages in novel view synthesis, novel pose rendering, and shape reconstruction.

Acknowledgement

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) through a Discovery Grant, as well as by Advanced Research Computing (ARC) at the University of British Columbia. Additional funding was provided in part by the Vector Institute for AI, the Canada CIFAR AI Chairs Program, and the NSERC Canada Research Chair (CRC) program. Partial support was also received from the Wallenberg AI, Autonomous Systems and Software Program (WASP), funded by the Knut and Alice Wallenberg Foundation, Sweden. We gratefully acknowledge ARC at UBC and Compute Canada for providing computational resources. We also thank the anonymous reviewers for their valuable feedback and insightful discussions.

References

[Aga18] AGARAP A. F.: Deep learning using rectified linear units (relu). *CoRR* (2018). 14

[BSAG18] BIŃKOWSKI M., SUTHERLAND D. J., ARBEL M., GRETTON A.: Demystifying MMD GANs. In *ICLR* (2018). 6

[BTS*16] BOUSMALIS K., TRIGEORGIS G., SILBERMAN N., KRISHNAN D., ERHAN D.: Domain separation networks. In *NeurIPS* (2016). 1, 2

[CXG*22] CHEN A., XU Z., GEIGER A., YU J., SU H.: Tensorf: Tensorial radiance fields. In *ECCV* (2022). 9

[DFG*22] DONG J., FANG Q., GUO Y., PENG S., SHUAI Q., ZHOU X., BAO H.: Totalselfscan: Learning full-body avatars from self-portrait videos of faces, hands, and bodies. In *NeurIPS* (2022). 2

[DZY*21] DU Y., ZHANG Y., YU H.-X., TENENBAUM J. B., WU J.: Neural radiance flow for 4d view synthesis and video processing. In *ICCV* (2021). 2

[GCS*20] GENOVA K., COLE F., SUD A., SARNA A., FUNKHOUSER T.: Local Deep Implicit Functions for 3D Shape. In *CVPR* (2020). 2

[GCV*19] GENOVA K., COLE F., VLASIC D., SARNA A., FREEMAN W. T., FUNKHOUSER T.: Learning Shape Templates with Structured Implicit Functions. *ICCV* (2019). 2

[GJC*23] GUO C., JIANG T., CHEN X., SONG J., HILLIGES O.: Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In *CVPR* (2023). 1, 2, 3, 4, 5, 6, 14, 18, 19, 22

[GPX*23] GENG C., PENG S., XU Z., BAO H., ZHOU X.: Learning neural volumetric representations of dynamic humans in minutes. In *CVPR* (2023). 2

[GSKH21] GAO C., SARAF A., KOPF J., HUANG J.-B.: Dynamic view synthesis from dynamic monocular video. In *ICCV* (2021). 2

[HLBK18] HUANG X., LIU M.-Y., BELONGIE S., KAUTZ J.: Multimodal unsupervised image-to-image translation. In *ECCV* (2018). 2

[HLX*23] HABERMANN M., LIU L., XU W., PONS-MOLL G., ZOLLHOEFER M., THEOBALT C.: Hdhumans: A hybrid approach for high-fidelity digital humans. *Proceedings of the ACM on Computer Graphics and Interactive Techniques* (2023). 1

[HRU*17] HEUSEL M., RAMSAUER H., UNTERTHINER T., NESSLER B., HOCHREITER S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS* (2017). 6

[HS81] HORN B. K., SCHUNCK B. G.: Determining optical flow. *Artificial intelligence* (1981). 17

[HZZ*24] HU L., ZHANG H., ZHANG Y., ZHOU B., LIU B., ZHANG S., NIE L.: Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In *CVPR* (2024). 3

[IRG*23] IŞIK M., RÜNZ M., GEORGOPOULOS M., KHAKHULIN T., STARCK J., AGAPITO L., NIESSNER M.: Humanrf: High-fidelity neural radiance fields for humans in motion. *ACM TOG* (2023). 10

[JCSH23] JIANG T., CHEN X., SONG J., HILLIGES O.: Instantavatar: Learning avatars from monocular video in 60 seconds. In *CVPR* (2023). 2

[JYS*22] JIANG W., YI K. M., SAMEI G., TUZEL O., RANJAN A.: Neuman: Neural human radiance field from a single video. In *ECCV* (2022). 1

[KB14] KINGMA D. P., BA J.: Adam: A method for stochastic optimization. *ICLR* (2014). 14

[KCG*24] KOCABAS M., CHANG J.-H. R., GABRIEL J., TUZEL O., RANJAN A.: Hugs: Human gaussian splats. *CVPR* (2024). 1, 3

[KKLD23] KERBL B., KOPANAS G., LEIMKÜHLER T., DRETTAKIS G.: 3d gaussian splatting for real-time radiance field rendering. *ACM TOG* (2023). 3

[KLF*23] KWON Y., LIU L., FUCHS H., HABERMANN M., THEOBALT C.: Deliffas: Deformable light fields for fast avatar synthesis. 1

[KMR*23] KIRILLOV A., MINTUN E., RAVI N., MAO H., ROLLAND C., GUSTAFSON L., XIAO T., WHITEHEAD S., BERG A. C., LO W.-Y., DOLLAR P., GIRSHICK R.: Segment anything. In *ICCV* (2023). 6

[LHR*21] LIU L., HABERMANN M., RUDNEV V., SARKAR K., GU J., THEOBALT C.: Neural actor: Neural free-view synthesis of human actors with pose control. *ACM TOG* (2021). 2

[LMTL21] LIN C.-H., MA W.-C., TORRALBA A., LUCEY S.: Barf: Bundle-adjusting neural radiance fields. In *ICCV* (2021). 2

[LNSW21] LI Z., NIKLAUS S., SNAVELY N., WANG O.: Neural scene flow fields for space-time view synthesis of dynamic scenes. In *CVPR* (2021). 2

[LSZ*22] LI T., SLAVCHEVA M., ZOLLHOEFER M., GREEN S., LASSNER C., KIM C., SCHMIDT T., LOVEGROVE S., GOESELE M., NEWCOMBE R., ET AL.: Neural 3d video synthesis from multi-view video. In *CVPR* (2022). 2

[LTH*18] LEE H.-Y., TSENG H.-Y., HUANG J.-B., SINGH M., YANG M.-H.: Diverse image-to-image translation via disentangled representations. In *ECCV* (2018). 2

[LTV*22] LI R., TANKE J., VO M., ZOLLHÖFER M., GALL J., KANAZAWA A., LASSNER C.: Tava: Template-free animatable volumetric actors. In *ECCV* (2022). 1, 2

[LVVPW22] LINDELL D. B., VAN VEEN D., PARK J. J., WETZSTEIN G.: BACON: Band-limited Coordinate Networks for Multiscale Scene Representation. *CVPR* (2022). 2

[LWP*24] LEI J., WANG Y., PAVLAKOS G., LIU L., DANIILIDIS K.: Gart: Gaussian articulated template models. *CVPR* (2024). 1, 3

[MESK22] MÜLLER T., EVANS A., SCHIED C., KELLER A.: Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *SIGGRAPH* (2022). 2, 9

[MON*19] MESCHEDER L., OECHSLE M., NIEMEYER M., NOWOZIN S., GEIGER A.: Occupancy Networks: Learning 3D Reconstruction in Function Space. In *CVPR* (2019). 2

[MSD*24] MOREAU A., SONG J., DHAMO H., SHAW R., ZHOU Y., PÉREZ-PELLITERO E.: Human gaussian splatting: Real-time rendering of animatable avatars. In *CVPR* (2024). 1, 3

[MST*20] MILDENHALL B., SRINIVASAN P. P., TANCIK M., BARRON J. T., RAMAMOORTHI R., NG R.: Nerf: Representing scenes as neural radiance fields for view synthesis. *ECCV* (2020). 2, 4

[NSLH21] NOGUCHI A., SUN X., LIN S., HARADA T.: Neural articulated radiance field. *ICCV* (2021). 1, 2

[OMT*21] OST J., MANNAN F., THUEREY N., KNODT J., HEIDE F.: Neural scene graphs for dynamic scenes. In *CVPR* (2021). 2

[PDW*21] PENG S., DONG J., WANG Q., ZHANG S., SHUAI Q., ZHOU X., BAO H.: Animatable neural radiance fields for modeling dynamic human bodies. In *ICCV* (2021). 2

[PFS*19] PARK J. J., FLORENCE P., STRAUB J., NEWCOMBE R., LOVEGROVE S.: DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation. In *CVPR* (2019). 2

[PGM*19] PASZKE A., GROSS S., MASSA F., LERER A., BRADBURY J., CHANAN G., KILLEEN T., LIN Z., GIMELSHEIN N., ANTIGA L., ET AL.: Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS* (2019). 14

[PSB*21] PARK K., SINHA U., BARRON J. T., BOUAZIZ S., GOLDMAN D. B., SEITZ S. M., MARTIN-BRUALLA R.: Nerfies: Deformable neural radiance fields. In *ICCV* (2021). 2

[PSH*21] PARK K., SINHA U., HEDMAN P., BARRON J. T., BOUAZIZ S., GOLDMAN D. B., MARTIN-BRUALLA R., SEITZ S. M.: Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *ACM TOG* (2021). 2

[PSSS19] PELEG T., SZEKELY P., SABO D., SENDIK O.: Im-net for high resolution video frame interpolation. In *CVPR* (2019). 2

[PZK*24] PANG H., ZHU H., KORTYLEWSKI A., THEOBALT C., HABERMANN M.: Ash: Animatable gaussian splats for efficient and photoreal human rendering. In *CVPR* (2024). 1

[PZX*21] PENG S., ZHANG Y., XU Y., WANG Q., SHUAI Q., BAO H., ZHOU X.: Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR* (2021). 2, 6, 14, 18, 19, 20, 22

[QWM*24] QIAN Z., WANG S., MIHAJLOVIC M., GEIGER A., TANG S.: 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. *CVPR* (2024). 1, 3, 6, 18, 19

[RLN*23] RHO D., LEE B., NAM S., LEE J. C., KO J. H., PARK E.: Masked wavelet representation for compact neural radiance fields. In *CVPR* (2023). 2

[RRC*16] RHODIN H., ROBERTINI N., CASAS D., RICHARDT C., SEIDEL H.-P., THEOBALT C.: General automatic human shape and motion capture using volumetric contour cues. In *ECCV* (2016). 1

[RRR*15] RHODIN H., ROBERTINI N., RICHARDT C., SEIDEL H.-P., THEOBALT C.: A versatile scene model with differentiable visibility applied to generative pose estimation. In *ICCV* (2015). 1

[SBR22] SU S.-Y., BAGAUTDINOV T., RHODIN H.: Danbo: Disentangled articulated neural body representations via graph neural networks. In *ECCV* (2022). 1, 2, 5, 14

[SBR23] SU S.-Y., BAGAUTDINOV T., RHODIN H.: Npc: Neural point characters from video. In *ICCV* (2023). 1, 2, 6, 7, 18, 19

[SEUD10] SALZMANN M., EK C. H., URTASUN R., DARRELL T.: Factorized orthogonal latent spaces. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (2010). 1, 2

[SLFB22] SHEKARFOROUSH S., LINDELL D., FLEET D. J., BRUBAKER M. A.: Residual multiplicative filter networks for multiscale reconstruction. *NeurIPS* (2022). 2

[SSC22] SUN C., SUN M., CHEN H.: Direct Voxel Grid Optimization: Super-fast Convergence for Radiance Fields Reconstruction. In *CVPR* (2022). 2

[SWR24] SONG C., WANDT B., RHODIN H.: Pose modulated avatars from video. In *ICLR* (2024). 1, 3, 5, 6, 18, 19

[SWS*25] SONG C., WU Z., SU S.-Y., WANDT B., SIGAL L., RHODIN H.: Locality sensitive avatars from video. In *ICLR* (2025). 3

[SWZ*19] SONG C., WU Z., ZHOU Y., GONG M., HUANG H.: Etnet: Error transition network for arbitrary style transfer. In *NeurIPS* (2019). 1, 2

[SYZR21] SU S.-Y., YU F., ZOLLHÖFER M., RHODIN H.: A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. *NeurIPS* (2021). 1, 2, 5, 14

[SZP*23] SONG C., ZHANG Y., PENG W., MOHAGHEGH P., WANDT B., RHODIN H.: Audioviewer: Learning to visualize sounds. 2

[TET*22] TAKIKAWA T., EVANS A., TREMBLAY J., MÜLLER T., MCGUIRE M., JACOBSON A., FIDLER S.: Variable bitrate neural fields. In *SIGGRAPH* (2022). 2

[TLY*21] TAKIKAWA T., LITALIEN J., YIN K., KREIS K., LOOP C., NOWROUZEZAHRAI D., JACOBSON A., MCGUIRE M., FIDLER S.: Neural Geometric Level of Detail: Real-time Rendering with Implicit 3D Shapes. In *CVPR* (2021). 2

[TTG*21] TRETSCHK E., TEWARI A., GOLYANIK V., ZOLLHÖFER M., LASSNER C., THEOBALT C.: Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *ICCV* (2021). 2

[WBSS04] WANG Z., BOVIK A. C., SHEIKH H. R., SIMONCELLI E. P.: Image quality assessment: from error visibility to structural similarity. *IEEE TIP* (2004). 6

[WCS*22] WENG C.-Y., CURLESS B., SRINIVASAN P. P., BARRON J. T., KEMELMACHER-SHLIZERMAN I.: Humannerf: Free-viewpoint rendering of moving people from monocular video. In *CVPR* (2022). 1, 3, 4, 6, 14, 18, 19

[WCT*24] WANG X., CHEN H., TANG S., WU Z., ZHU W.: Disentangled representation learning. *IEEE TPAMI* (2024). 2

[WJMY23] WU Z., JIN Y., MOO YI K.: Neural fourier filter bank. In *CVPR* (2023). 2, 9

[WLL*21] WANG P., LIU L., LIU Y., THEOBALT C., KOMURA T., WANG W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *NeurIPS* (2021). 2

[WSC*22] WU Z., SONG C., CHEN G., GUO S., HUANG W.: Completeness and coherence learning for fast arbitrary style transfer. *Transactions on Machine Learning Research* (2022). 1, 2

[WSGT22] WANG S., SCHWARZ K., GEIGER A., TANG S.: Arah: Animatable volume rendering of articulated human sdfs. In *ECCV* (2022). 1, 2, 5, 6, 7, 20

[WSW22] WANG Y., SKOROKHODOV I., WONKA P.: Hf-neus: Improved surface reconstruction using high-frequency details. *NeurIPS* (2022). 2

[WSZ*20] WU Z., SONG C., ZHOU Y., GONG M., HUANG H.: Efanet: Exchangeable feature alignment network for arbitrary style transfer. In *AAAI* (2020). 1, 2

[WWL*19] WU Z., WANG X., LIN D., LISCHINSKI D., COHEN-OR D., HUANG H.: Sagnet: Structure-aware generative network for 3d-shape modeling. *ACM TOG* (2019). 2

[WZR*24] WEN J., ZHAO X., REN Z., SCHWING A. G., WANG S.: Gomavatar: Efficient animatable human modeling from monocular video using gaussians-on-mesh. In *CVPR* (2024). 6, 18, 19

[XAS21] XU H., ALLDIECK T., SMINCHISESCU C.: H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. *NeurIPS* (2021). 2

[XCZ*18] XU W., CHATTERJEE A., ZOLLHÖFER M., RHODIN H., MEHTA D., SEIDEL H.-P., THEOBALT C.: Monoperfcap: Human performance capture from monocular video. *ACM TOG* (2018). 1

[XTS*22] XIE Y., TAKIKAWA T., SAITO S., LITANY O., YAN S., KHAN N., TOMBARI F., TOMPKIN J., SITZMANN V., SRIDHAR S.: Neural Fields in Visual Computing and Beyond. *Computer Graphics Forum* (2022). 2

[YCFB*21] YEN-CHEN L., FLORENCE P., BARRON J. T., RODRIGUEZ A., ISOLA P., LIN T.-Y.: inerf: Inverting neural radiance fields for pose estimation. In *IROS* (2021). 2

[YCL*23] YU Z., CHENG W., LIU X., WU W., LIN K.-Y.: Monohuman: Animatable human neural field from monocular video. In *CVPR* (2023). 1, 3, 6, 14, 18, 19

[YGKL21] YARIV L., GU J., KASTEN Y., LIPMAN Y.: Volume rendering of neural implicit surfaces. In *NeurIPS* (2021). 2, 5, 14

[YLSL21] YUAN W., LV Z., SCHMIDT T., LOVEGROVE S.: Star: Self-supervised tracking and reconstruction of rigid objects in motion with neural rendering. In *CVPR* (2021). 2

[YYTK21] YU A., YE V., TANCIK M., KANAZAWA A.: pixelnerf: Neural radiance fields from one or few images. In *CVPR* (2021). 2

[ZHY*22] ZHENG Z., HUANG H., YU T., ZHANG H., GUO Y., LIU Y.: Structured local radiance fields for human avatar modeling. In *CVPR* (2022). 2

[ZIE*18] ZHANG R., ISOLA P., EFROS A. A., SHECHTMAN E., WANG O.: The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR* (2018). 6

[ZZZ*23] ZHENG Z., ZHAO X., ZHANG H., LIU B., LIU Y.: Avatarrex: Real-time expressive full-body avatars. *ACM TOG* (2023). 2