Expressive Whole-Body 3D Gaussian Avatar

Gyeongsik Moon[1,2], Takaaki Shiratori[2], and Shunsuke Saito[2]

[1]DGIST
mks0601@dgist.ac.kr

[2]Codec Avatars Lab, Meta
{tshiratori,shunsukesaito}@meta.com

https://mks0601.github.io/ExAvatar

Abstract. Facial expression and hand motions are necessary to express our emotions and interact with the world. Nevertheless, most of the 3D human avatars modeled from a casually captured video only support body motions without facial expressions and hand motions. In this work, we present **ExAvatar**, an expressive whole-body 3D human avatar learned from a short monocular video. We design ExAvatar as a combination of the whole-body parametric mesh model (SMPL-X) and 3D Gaussian Splatting (3DGS). The main challenges are 1) a limited diversity of facial expressions and poses in the video and 2) the absence of 3D observations, such as 3D scans and RGBD images. The limited diversity in the video makes animations with novel facial expressions and poses non-trivial. In addition, the absence of 3D observations could cause significant ambiguity in human parts that are not observed in the video, which can result in noticeable artifacts under novel motions. To address them, we introduce our hybrid representation of the mesh and 3D Gaussians. Our hybrid representation treats each 3D Gaussian as a vertex on the surface with pre-defined connectivity information (*i.e.*, triangle faces) between them following the mesh topology of SMPL-X. It makes our ExAvatar animatable with novel facial expressions by driven by the facial expression space of SMPL-X. In addition, by using connectivity-based regularizers, we significantly reduce artifacts in novel facial expressions and poses.

1 Introduction

Humans use all facial expressions, body motions, and hand motions to express our emotions and intentions, and interact with other people and objects. In particular, facial expressions and hand gestures are one of the most powerful channels for non-verbal communication, and hand motions are necessary to interact with diverse types of objects. Modeling the facial expression, body motion, and hand motion altogether is extremely challenging. Several whole-body 3D human geometry models have been introduced [2,18,32,43]. Among them, SMPL-X [32] is the most widely used one, which motivated a number of 3D whole-body pose estimation methods [4, 9, 11, 22, 24, 28, 39, 44] and benchmarks [31].

To represent 3D humans beyond the minimally clothed parametric models, personalized 3D human avatars have been recently studied. The 3D human avatar is a representation that combines 3D geometry and the appearance of a

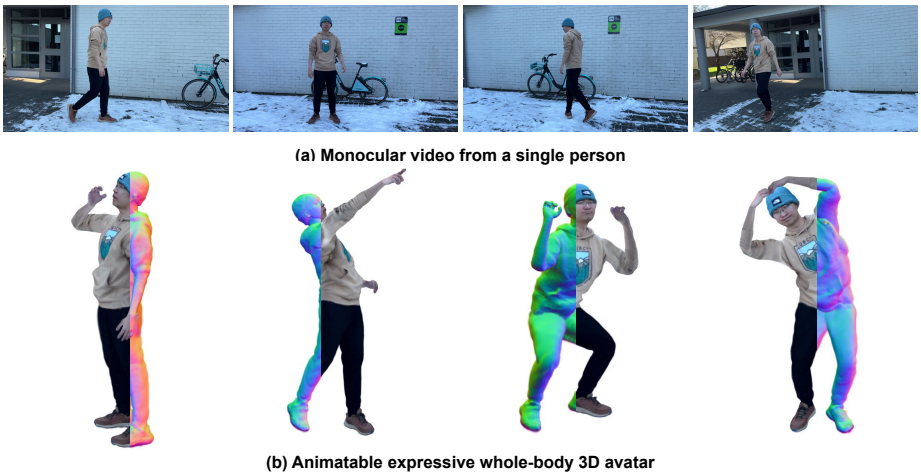

Fig. 1: From (a) a monocular video from a single person, we create our (b) **ExAvatar**, an expressive whole-body 3D avatar, animatable with novel facial expression code, hand poses, and body poses of SMPL-X.

certain person, which can be animated and rendered with novel poses. However, most of existing 3D human avatars [6, 8, 13, 15–17, 20, 21, 33, 34] modeled from a casually captured video only support body motions without facial expressions and hand motions. Their avatars bake facial expressions and hand poses, and animating them is not possible. A recent work [41] introduced a whole-body avatar that supports animation with facial expressions, and body and hand poses; however, it requires 3D observations, such as 3D scans or RGBD images with highly accurate SMPL-X registrations, with diverse poses and facial expressions. Such an assumption does not hold for the majority of casually captured videos in daily life.

We present **ExAvatar**, an expressive whole-body 3D human avatar that can be made from a short monocular video. ExAvatar is designed as a combination of the whole-body 3D parametric model (SMPL-X) [32] and 3D Gaussian Splatting (3DGS) [19]. It utilizes the whole-body drivability of SMPL-X and the photorealistic and efficient rendering capability of 3DGS. After the training, it is animatable with novel facial expression code and 3D pose of SMPL-X, as shown in Fig. 1. Despite its desired properties, modeling ExAvatar is an non-trivial task with the following two challenges: 1) a limited diversity of facial expressions and poses in the video and 2) the absence of 3D observations, such as 3D scans and RGBD videos. The limited diversity in the video makes a drivability with novel facial expressions and poses non-trivial. In addition, the absence of 3D observations creates ambiguity in the occluded human parts, exhibiting noticeable artifacts in novel facial expressions and poses.

To address them, we propose a novel hybrid representation of the surface mesh and 3D Gaussians in ExAvatar. Our hybrid representation treats each 3D Gaussian as a vertex on the surface, where the vertices have pre-defined connectivity (*i.e.*, triangle faces) between them following the mesh topology of

SMPL-X. Existing volumetric avatars [6,8,13,16,17,21,33,34,41] do not have the connectivity by the definition. Also, previous 3DGS-based [15,20] works consider a set of 3D Gaussian points as a point cloud without considering the connectivity between them.

Using our hybrid representation, our ExAvatar becomes fully compatible with the facial expression space of SMPL-X. Therefore, it can be driven with any facial expression code of SMPL-X *even from a short monocular video without diverse facial expressions*. As our 3D Gaussians share the exactly same mesh topology with SMPL-X, we simply add the vertex offsets to our 3D Gaussian points to move them according to the facial expression as in FLAME [23] and SMPL-X [32]. Hence, unlike previous works [41], our drivability of the facial expression is not strictly limited by the number of training frames (*e.g.*, 30 seconds of a short video).

Another benefit is that we can significantly reduce artifacts in novel facial expressions and poses using connectivity-based regularizers. As the pose diversity in the training set is very limited, there could be human parts that are not observed at all in the video. Without 3D observations, the ambiguous human parts could suffer from artifacts in novel poses. While several point-based regularizers (e.g., L2 regularization of the underlying SMPL/SMPL-X template mesh) have been proposed [15], they do not consider *connectivity* between vertices. Such a lack of connectivity could introduce floating 3D Gaussians.By considering the connectivity, we can naturally enforce local similarity, significantly reducing artifacts.

Throughout our experiments, our method substantially outperforms all previous 3D human avatars in various benchmarks. Our contributions can be summarized as follows.

- We present ExAvatar, an expressive whole-body 3D human avatar that can be made from a short monocular video without requiring 3D observations.
- We propose a hybrid representation of the surface mesh and 3D Gaussians. It allows ExAvatar to be animated with any novel facial expression code of SMPL-X even from a short monocular video without diverse facial expressions.
- Using connectivity information between 3D Gaussians, we significantly reduce artifacts, especially in novel facial expressions and poses.

2 Related works

3D human avatars. Various 3D representations are used for modeling 3D human avatars. Alldieck *et al*. [1] extended SMPL mesh with per-vertex offsets. Bagautdinov *et al*. [3] achieved high-fidelity results using conditional variational autoencoder, which can be animated with incomplete driving signals. Remelli *et al*. [38] propose to use texel-aligned features, a localized representation. Motivated by neural radiance fields [27], many volumetric and implicit representation-based avatars have been introduced. Peng *et al*. [33, 34] created a 3D human avatar from a capture studio, which provides accurate 3D pose and

multi-view images for supervision. Kwon *et al*. [21] improved the previous works by utilizing vertex-aligned features. Shen *et al*. [41] created a whole-body 3D avatar, which supports whole-body animation with facial expression, from their capture studio dataset. In contrast to the above works that make 3D avatars from capture studios, recent works focus on making 3D avatars from a short monocular video without requiring 3D observations, such as 3D scans, RGBD images, or multi-view images. Jiang *et al*. [17] introduced a dataset and method for making a 3D human avatar from a short monocular video taken from in-the-wild environments. Guo *et al*. [13] proposed a system that can decompose a scene and human with self-supervised learning. Jiang *et al*. [16] introduced a system that can make a 3D human within several minutes. Recently introduced 3DGS [19], which achieves both powerful and efficient rendering capability, motivated several 3DGS-based avatars [15,20,26]. Kocabas *et al*. [20] use the triplane for creating 3D avatar. Hu *et al*. [15] introduced a robust system that takes a positional map of a posed SMPL mesh. Moon *et al*. [30] presented universal hand model (UHM) to create authentic hand avatars from a phone scan. Chen *et al*. [7] extended UHM of Moon *et al*. [30] for the relightability.

Except for a few works [3, 26, 41], most of the above works only support body motions without hand motions and facial expressions. X-Avatar [41] supports whole-body animation including facial expressions; however, it has two limitations. First, it requires diverse facial expressions with accurate 3D geometry registration in videos to create avatars. This is because they cannot directly utilize the facial expression space of FLAME [23]/SMPL-X [32]. They need to transform the mesh-based facial expression space of FLAME [23]/SMPL-X [32] to their implicit representation using learnable modules. To train the transformation module, they need training data with sufficiently diverse facial expressions and accurate 3D geometry registrations. Second, it requires 3D observations, such as 3D scans or RGBD images with accurate SMPL-X registrations, for the training, hard to obtain from in-the-wild environments. Due to the above two reasons, X-Avatar [41] is hard to apply to practical settings, such as short monocular videos. Liu *et al*. [26] proposed another whole-body 3D avatar; however, their avatar is not animated with novel facial expressions.

Whole-body 3D human modeling and perception. Modeling face, body, and hands at the same time is an extremely challenging problem as each human part has its own different characteristics. Several whole-body 3D human models have been introduced [2, 18, 32, 43], which model 3D geometry of minimally clothed humans. They are parametric models, parameterized by 3D poses, facial expression code, and shape parameter. Among them, SMPL-X [32] is the most widely used one due to its completeness. Motivated by the optimization baseline [32] and benchmarks [31], a number of 3D whole-body pose estimation methods [4, 9, 11, 22, 24, 28, 39, 44] have been introduced.

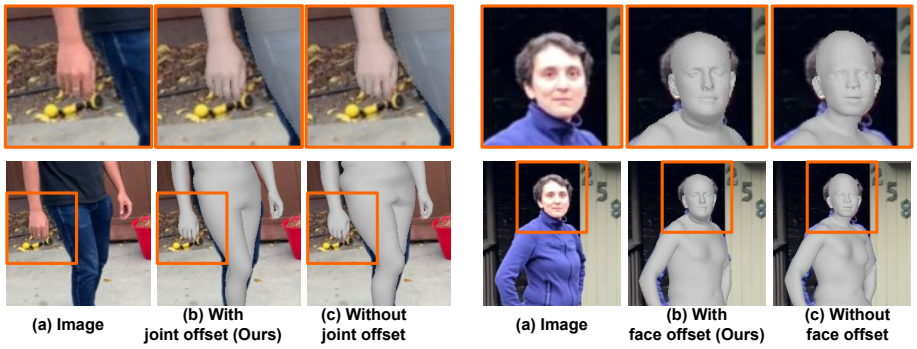

Fig. 2: The effectiveness of our joint offset $\Delta\mathbf{J}$ and face offset $\Delta\mathbf{V}_{\text{face}}$. They are necessary for the accurate registration of hands and face, which results in accurate co-registration of the whole body.

3 ExAvatar

3.1 Accurate co-registration of SMPL-X

We assume videos, which usually consist of 30 seconds of frames, are taken from an in-the-wild environment. The video is from a single person with a natural backgrounds. Before training our ExAvatar, we first preprocess the video. Following previous works [13,17], we first run an off-the-shelf SMPL-X regressor [22] and 2D pose estimator [10] to all frames. Then, we additionally fit the regressed SMPL-X parameters (*i.e.*, 3D poses $\theta \in \mathbb{R}^{55\times3}$, shape parameter $\beta \in \mathbb{R}^{100}$, and facial expression code $\psi \in \mathbb{R}^{50}$) and 3D translation $\mathbf{t}$ to the estimated 2D pose of each frame. The shape parameter is shared across all frames as all frames are from the same person.

One challenge during registering SMPL-X parameters to a video is accurate *co-registration* of body, hands, and face, unique challenges of the whole-body avatar. The registration of hands and face can be negatively affected by a limited expressiveness of SMPL-X and registration accuracy of the body, which can limit the co-registration accuracy. To achieve the *accurate co-registration* of body, hands, and face, we introduce two optimizable offsets initialized with zero and shared across all frames. Both offsets are identity (ID)-dependent offsets and are not dependent on the poses and facial expressions. Hence, they are added to the T-pose template mesh of SMPL-X before performing the linear blend skinning (LBS).

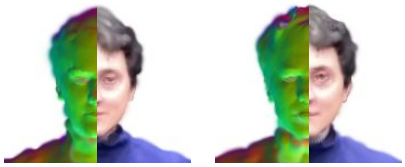

Fig. 3: Without the face offset $\Delta\mathbf{V}_{\text{face}}$, the final 3D geometry of the avatar becomes totally inauthentic and inaccurate. For each setting, normals of 3D Gaussian points and colors are used for the rendering.

First, we introduce joint offset $\Delta\mathbf{J}$, added to the joints in the T-pose space of SMPL-X. The joint offset $\Delta\mathbf{J}$, which affect both 3D skeleton and surface, are

especially helpful to fit hands more perfectly as the shape parameter of SMPL-X has limited coverage of 3D hand skeleton, as shown in Fig. 2 left. Second, we introduce face offset $\Delta\mathbf{V}_{\text{face}}$, per-vertex offset of the face region of SMPL-X, added to the face vertex in the T-pose space of SMPL-X. To optimize the face offset $\Delta\mathbf{V}_{\text{face}}$, we first fit the 3D face-only model (*i.e.*, FLAME [23]) to 2D poses and images of the face by running DECA [12] and further optimizing it to 2D poses. Then, we optimize the face offset by making a summation of the face offset and 3D face vertices of SMPL-X close to fitted FLAME vertices. The optimization is straightforward as the face region of SMPL-X has exactly the same topology as that of FLAME. The rationale is that 1) the face-only model has higher expressiveness in its shape space than the whole-body model and 2) the registration of the face-only model is not affected by the body registration. Fig. 2 right and 3 show the effectiveness of our face offsets. Such a special treatment in the registration stage is greatly helpful for the final 3D avatar, not considered in previous whole-body avatars [26, 41]. Please refer to the supplementary material about the details of the fitting.

3.2 Architecture

Fig. 4 shows the architecture of ExAvatar. We model ExAvatar on top of a canonical 3D human mesh, denoted by $\bar{\mathbf{V}} \in \mathbb{R}^{N\times 3}$, where it has $N = 167$K upsampled vertices and 335K upsampled triangle faces. To obtain it, we first pass the optimized SMPL-X shape parameter β, joint offsets $\Delta\mathbf{J}$, and face offsets $\Delta\mathbf{V}_{\text{face}}$ from Sec. 3.1, and a pre-defined neural pose (*i.e.*, 大-pose) to the SMPL-X layer. Then, we upsample it with the subdivision function of PyTorch3D [37], which can upsample other 3D assets, such as facial expression blend shapes, in a consistent way.

Per-vertex Gaussian assets regression. We initialize a learnable triplane [5] $\mathbf{T} \in \mathbb{R}^{3\times C\times H\times W}$ with zero, where $C = 32$, $H = 128$, and $W = 128$ represent channel dimension, height, and width of the triplane, respectively. Then, we prepare a positional encoding mesh $\bar{\mathbf{P}} \in \mathbb{R}^{N\times 3}$ with a pre-defined neutral pose (*i.e.*, 大-pose) and zero shape parameter. We upsample the positional encoding mesh with the above subdivision function, which produces the same mesh topology as the canonical mesh $\bar{\mathbf{V}}$. We extract the per-vertex feature from the triplane by orthogonally projecting $\bar{\mathbf{P}}$ to each plane and performing the bilinear interpolation. The triplane is useful as it naturally enforces similarity between close vertices. In practice, we construct another triplane dedicated to the face, as the face requires detailed geometry and appearance modeling with a small physical size. The reason for not using the canonical mesh $\bar{\mathbf{V}}$ for the feature extraction is that it keeps changing during the training as we further optimize the shape parameter β and the joint offset $\Delta\mathbf{J}$ during the training. If the position of a certain vertex changes, the extracted triplane feature of that vertex could be one that was from other vertices, which can make the training unstable.

The interpolated features from the triplane are concatenated, denoted by $\mathbf{F} \in \mathbb{R}^{N\times 96}$. We pass $\mathbf{F}$ to two multi-layer perceptrons (MLPs), which regress 1) 3D offset $\Delta\mathbf{V}_{\text{tri}} \in \mathbb{R}^{N\times 3}$ and scale $\mathbf{S}_{\text{tri}} \in \mathbb{R}^{N\times 1}$ and 2) RGB values $\mathbf{C}_{\text{tri}} \in \mathbb{R}^{N\times 3}$

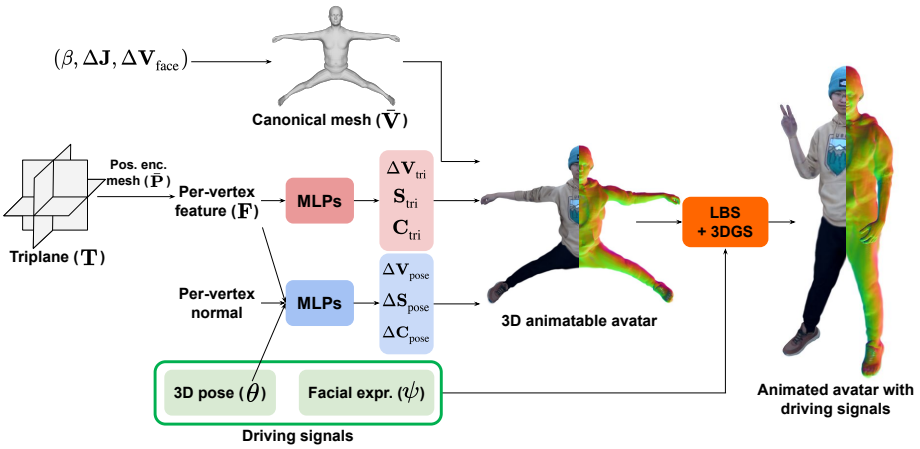

Fig. 4: The architecture of our ExAvatar. From the canonical mesh $\bar{\mathbf{V}}$, triplane $\mathbf{T}$, per-vertex normal, and 3D pose θ, we build a 3D animatable avatar. Then, with driving signals, 3D pose θ and facial expression code ψ of SMPL-X [32], we animate the avatar and render it to the screen space with 3DGS [19]. For the normal rendering, we calculate the normal vectors using the positions of 3D Gaussian points and mesh topology of SMPL-X.

for the 3DGS, respectively. The MLPs are shared across all vertices. Motivated by Hu *et al*. [15], for better generalization to novel viewpoints, we limit all Gaussian assets to be isotropic by limiting a degree of freedom of the scale to 1 and setting the rotation and opacity to identity and one, respectively. Please refer to the supplementary material for the detailed architecture of the MLPs. The regressed Gaussian assets (*i.e.*, 3D offset, scale, and RGB values) are solely from the triplane, shared across all frames. Hence, they represent identity (ID)-dependent and environment (*e.g.*, lighting)-dependent assets without pose dependency as ID and environment are fixed in the input video, while pose changes for each frame.

To additionally model pose-dependent deformations, we employ two additional MLPs. The first MLP takes $\mathbf{F}$ and 3D poses θ without the root pose and outputs 3D vertex offset $\Delta\mathbf{V}_{\text{pose}} \in \mathbb{R}^{N\times 3}$ and scale offset $\Delta\mathbf{S}_{\text{pose}} \in \mathbb{R}^{N\times 1}$. The second MLP takes $\mathbf{F}$, 3D poses θ without the root pose, and the normal vector of each vertex and outputs RGB offset $\Delta\mathbf{C}_{\text{pose}} \in \mathbb{R}^{N\times 3}$. The additional normal vector can 1) provide the view-dependent shading information to the network and 2) be useful to disentangle geometry and appearances [13, 40]. Thanks to our hybrid representation, we can easily obtain the per-vertex normal vector by averaging normals of triangle faces that include the vertex. Instead of directly predicting pose-dependent Gaussian assets, ours output pose-dependent offsets. This is helpful for the generalization to novel poses as Gaussian assets solely from the triplane already have reasonable expressiveness, which makes the role of the pose-dependent Gaussian assets small. Such a design is especially important when making 3D avatars from a short video like ours as limited pose diversity makes generalization to novel poses challenging.

3.3 Animation and rendering

Fig. 5 shows examples of our animated and rendered avatars, made from short monocular videos.

Animation. We need to animate Gaussian points from the canonical space with given facial expression code ψ and 3D poses θ of SMPL-X. To this end, we first replace the pose-dependent vertex offset $\varDelta\mathbf{V}_{\text{pose}}$ of hand and face vertices to those of SMPL-X. This is because the hand and face are often naked; hence, we can directly utilize vertex offsets of SMPL-X. Then, we add vertex offsets from facial expression code ψ of SMPL-X to face vertices. *By directly using the facial expression offsets of SMPL-X, we do not have to learn a new facial expression space.* Such a direct utilizing is from our *hybrid representation of the mesh and 3D Gaussians*. The below equations describe the above deformations in the *canonical space*.

$$\bar{\mathbf{V}}_{\text{tri}} = \bar{\mathbf{V}} + \varDelta\mathbf{V}_{\text{tri}} + \varDelta\mathbf{V}_{\text{expr}}, \tag{1}$$

$$\bar{\mathbf{V}}_{\text{pose}} = \bar{\mathbf{V}} + \varDelta\mathbf{V}_{\text{tri}} + \varDelta\mathbf{V}_{\text{pose}} + \varDelta\mathbf{V}_{\text{expr}}, \tag{2}$$

where $\varDelta\mathbf{V}_{\text{expr}}$ represents the facial expression offset of SMPL-X, obtained from the facial expression code ψ. Then, for the body vertices, we take the skinning weight of the nearest vertices from downsampled $\bar{\mathbf{V}}$, while for the hand and face vertices, we use the original skinning weight of the vertices. This is because, for the body vertices, their semantic meaning could change due to the cloth geometry. The final animated geometry, $\mathbf{V}_{\text{tri}}$ and $\mathbf{V}_{\text{pose}}$, are represented with below equations.

$$\mathbf{V}_{\text{tri}} = \text{LBS}(\bar{\mathbf{V}}_{\text{tri}}, \theta, \mathbf{W}_{\text{tri}}) \quad \text{and} \quad \mathbf{V}_{\text{pose}} = \text{LBS}(\bar{\mathbf{V}}_{\text{pose}}, \theta, \mathbf{W}_{\text{pose}}), \tag{3}$$

where $\mathbf{W}_{\text{tri}}$ and $\mathbf{W}_{\text{pose}}$ represent the skinning weight of $\bar{\mathbf{V}}_{\text{tri}}$ and $\bar{\mathbf{V}}_{\text{pose}}$, respectively.

Rendering. To render animated 3D geometry, we use 3DGS rendering pipeline [19] like the below equations.

$$\mathbf{I}_{\text{tri}} = f(\mathbf{V}_{\text{tri}}, \exp(\mathbf{S}_{\text{tri}}), \mathbf{C}_{\text{tri}}, \mathbf{K}, \mathbf{E}), \tag{4}$$

$$\mathbf{I}_{\text{pose}} = f(\mathbf{V}_{\text{pose}}, \exp(\mathbf{S}_{\text{tri}} + \varDelta\mathbf{S}_{\text{pose}}), \mathbf{C}_{\text{tri}} + \varDelta\mathbf{C}_{\text{pose}}, \mathbf{K}, \mathbf{E}), \tag{5}$$

where f, $\mathbf{K}$, and $\mathbf{E}$ represent rendering function of 3DGS, camera intrinsic, and extrinsic matrices, respectively. As described above, following Hu *et al.* [15], we restrict all Gaussian assets to isotropic for better generalization; hence, rotation and opacity of all Gaussian points are set to identity and one, respectively, not described in the equations.

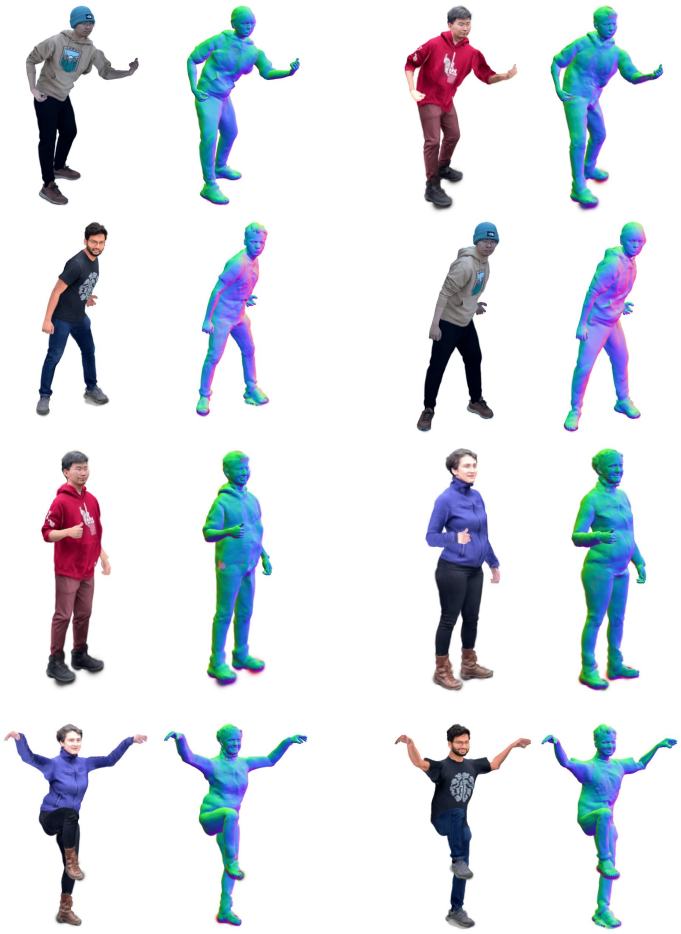

Fig. 5: Our animated expressive whole-body avatars, made from monocular videos of NeuMan dataset [17]. Avatars of each row are animated with the same facial expression code ψ and 3D pose θ of SMPL-X.

3.4 Loss functions

During the training of our ExAvatar, we optimize the triplane $\mathbf{T}$, MLPs for the regression of Gaussian assets in Sec. 3.2, 3D pose θ of each frame, facial expression code ψ of each frame, 3D translation $\mathbf{t}$ of each frame, the shape parameter β, and the joint offset $\Delta\mathbf{J}$. Also, we simultaneously optimize a 3DGS for the background following the original implementation [19] by segmenting out human regions using human masks from an off-the-shelf human segmentation model [14]. Modeling background simultaneously produces better foreground mask [13], as estimated masks often have errors, especially on hand parts. We denote rendered images from a combination of Eq. 4 and the 3DGS for the background by $\mathbf{I}^{*}_{\text{tri}}$. Likewise, we denote rendered images from a combination of

Fig. 6: The effectiveness of our face loss. Without the face loss, the geometry and texture of the face are not consistent, which makes significant artifacts when driving. The right one (b) shows that without the face loss, when the mouth is opened, the upper lip remains at the same position, while only below lip is opened.

Fig. 7: The effectiveness of the Laplacian regularizer, which makes 3D avatars in novel facial expressions and poses greatly stable. On the other hand, the widely used $L2$ regularizer to the distance from SMPL-X surface to 3D Gaussian points suffers from severe artifacts. We successfully incorporated the Laplacian regularizer using our hybrid representation of the surface mesh and 3D Gaussians.

Eq. 5 and the 3DGS for the background by $\mathbf{I}^*_{\text{pose}}$. To train ExAvatar, we minimize the below loss functions.

Image loss. Following 3DGS [19], we minimize $L1$ distance, 1 - SSIM, and LPIPS [45] between rendered images (*i.e.*, $\mathbf{I}^*_{\text{tri}}$ and $\mathbf{I}^*_{\text{pose}}$) and the captured image. We found that the additional LPIPS is helpful for sharper textures. To save the computation, we compute the image loss after cropping the human region.

Face loss. Unlike other human parts, the face has its unique characteristics as there should be a strong consistency between geometry and texture. For example, lip geometry usually has reddish textures. If other face geometry has lip textures, in novel facial expressions or jaw poses, the lip would not properly change, which can lead to significant artifacts, as shown in Fig. 6 (b). Simply minimizing the above image loss does not guarantee the consistency between the geometry and texture of the face region. To enforce the consistency, we minimize the $L1$ distance between the rendered face image with a standard differentiable mesh renderer and the captured image, where the texture for the mesh renderer is prepared by averaging the unwrapped UV texture of the face-only model [23] registrations from Sec. 3.1. The UV texture is fixed, and the positions of 3D Gaussian points of the face region are adjusted to minimize the loss function.

Fig. 6 (a) shows the effectiveness of our face loss functions. Thanks to our hybrid representation of the mesh and 3D Gaussians, such a mesh-based loss function can be easily incorporated into our system.

Regularizers. Due to the limited pose diversity in the training set, there can be human parts that are not observed in the video. Such human parts suffer from occlusion ambiguity, which could result in artifacts in novel facial expressions and poses. In addition, to utilize the facial expression offsets of SMPL-X, we need to make the face geometry similar to that of SMPL-X. To address them, we utilize connectivity-based regularizers (*i.e.*, Laplacian regularizer), motivated by the mesh modeling works [25, 29]. Fig. 7 shows that our connectivity-based regularizer significantly reduces artifacts in novel facial expressions and poses. We minimize the difference of the 1) Laplacian of deformed 3D Gaussian points in the canonical space (*i.e.*, $\bar{\mathbf{V}}_{\text{tri}}$ and $\bar{\mathbf{V}}_{\text{pose}}$) and 2) Laplacian of the canonical mesh $\bar{\mathbf{V}}$. In this way, we can easily encourage the local similarity between 3D Gaussian points, which can prevent floating 3D Gaussians.In particular, our connectivity-based regularizer is much more effective than the widely used $L2$ regularizer [15] that simply penalizes distance between 3D Gaussian points and underlying template mesh without considering the connectivity information. Due to our hybrid representation of the mesh and 3D Gaussians, the Laplacian regularizer, widely used in mesh modeling works, can be easily included in our system. In addition to regularizing the 3D positions of 3D Gaussian points, we compute the same Laplacian regularizer for the scales and RGBs of 3D Gaussian points. For other regularizers, please refer to the supplementary material.

4 Experiments

4.1 Datasets

NeuMan. NeuMan [17] provides several short monocular videos taken from in-the-wild environments. Each video contains a single person walking around for about 15 seconds. Following previous works [15] we use *bike, citron, jogging, and seattle* videos that exhibit most human body regions and contain minimal blurry images. We follow their official training and testing splits.

X-Humans. X-Humans [41] provides 3D scans and RGBD videos of multiple subjects, captured from a studio. Compared to NeuMan, X-Humans has more diverse facial expressions and hand poses. There are two experimental protocols: 1) using 3D scans and 2) using RGBD images for creating avatars. We create our avatar only with monocular RGB videos *without depthmaps* and compare ours against previous works [41] that use RGBD videos. We use *0028, 0034, and 0087* subjects as their pre-trained weights of the RGBD protocol are publicly available. We follow their official training and testing splits.

4.2 Comparison to state-of-the-art methods

Tab. 1 and 2 show that our ExAvatar achieves the best results on NeuMan [17] dataset regardless of whether the rendered background pixels are included or

Table 1: Comparisons of 3D human avatars on the test set of NeuMan [17] Rendered backgrounds are considered in the evaluation. Only our ExAvatar supports face and hand animations.

Methods	PSNR↑	SSIM↑	LPIPS↓
NeuMan [17]	24.22	0.77	0.27
Vid2Avatar [13]	15.41	0.53	0.66
HUGS [20]	25.17	0.83	0.16
ExAvatar (Ours)	**27.47**	**0.90**	**0.10**

Table 2: Comparisons of 3D human avatars on the test set of NeuMan [17]. Rendered backgrounds are **not** considered in the evaluation. Only our ExAvatar supports face and hand animations.

Methods	PSNR↑	SSIM↑	LPIPS↓
HumanNeRF [42]	27.06	0.967	0.019
InstantAvatar [16]	28.47	0.972	0.028
NeuMan [17]	29.32	0.972	0.014
Vid2Avatar [13]	30.70	0.980	0.014
GaussianAvatar [15]	29.94	0.980	0.012
3DGS-Avatar [36]	28.99	0.974	0.016
ExAvatar (Ours)	**34.80**	**0.984**	**0.009**

Table 3: Comparisons of 3D human avatars on the test set of X-Humans [41]. Methods with * use additional depth maps for the training.

Methods	*00028*			*00034*			*00087*		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
X-Avatar [41]*	28.57	0.976	0.026	28.05	0.965	0.035	30.89	0.970	0.030
ExAvatar (Ours)	**30.58**	**0.981**	**0.018**	**28.75**	**0.966**	**0.029**	**32.01**	**0.972**	**0.025**

not. All numbers are from papers [15, 20] except NeuMan [17], Vid2Avatar [13], and 3DGS-Avatar [36] of Tab. 2, measured with their officially released code. To exclude background pixels, we used an off-the-shelf segmentation network [14] following GaussianAvatar [15]. Following previous works [6, 15, 16, 36], for the evaluation on NeuMan dataset, we fit SMPL-X parameters of testing frames while freezing all other parameters with the image loss of Sec. 3.4.

Fig. 8 shows that ours produces photorealistic renderings in novel views and poses. For example, prints on the shirts (the first and third rows) are significantly sharper and clearer than those of previous works. Most importantly, ours produces faces and hands in novel views and poses substantially better than previous avatars. As previous avatars do not have controllability on faces and hands, the averaged blurry textures are baked in (faces in the first row and hands in the second row and fourth row). On the other hand, ours has sharp textures benefiting from the whole-body modeling.

Tab. 3 and Fig. 9 show that our ExAvatar outperforms previous whole-body avatar [41] on X-Humans [41] dataset even without using depth maps, while the previous work relies on it. Our hybrid representation of the surface mesh and 3D Gaussians leads to stable training and shaper textures of faces and hands. For X-Avatar's results, we used their officially released pre-trained weights and code. Following Shen *et al.* [41], we used given SMPL-X parameters without further fitting them to testing frames.

4.3 Ablation study

In this section, we ablate the effectiveness of our hybrid representation of the surface mesh and 3D Gaussians, which enables us to incorporate Laplacian regularizer and face loss into our system. Fig. 7 and Tab. 4 show that incorporating

Fig. 8: Qualitative comparison of our ExAvatar, Vid2Avatar [13], NeuMan [17], and 3DGS-Avatar [36] on the test set of Neuman [17].

Table 4: Ablation study for the effectiveness of incorporating Laplacian regularizer to our 3D Gaussian-based system on the test set of NeuMan [17].

Settings	PSNR↑	SSIM↑	LPIPS↓
Without Lap. reg.	28.21	0.968	0.199
With Lap. reg. (Ours)	**34.80**	**0.984**	**0.009**

Table 5: Ablation study for the effectiveness of our face loss on the cropped face images of a test set of X-Humans [41].

Settings	PSNR↑	SSIM↑	LPIPS↓
Without face loss	20.02	0.671	**0.06**
With face loss (Ours)	**22.07**	**0.693**	**0.06**

the Laplacian regularizer into our system brings significant performance boost and stability. Fig. 6 and Tab. 5 show the benefit of the proposed face loss. The numbers in Tab. 5 are measured only for cropped face images when the face is visible to evaluate the effectiveness of the face loss. The face visibility is decided by rasterizing SMPL-X meshes and checking the number of rasterized triangle faces of the face region.

5 Conclusion

Summary. We present **ExAvatar**, an expressive whole-body 3D avatar that can be made from a short monocular video. We propose a hybrid representation of the surface mesh and 3D Gaussians to address 1) the limited diversity of facial expressions and poses in the video and 2) the absence of 3D observations, such as 3D scans and RGBD images. Our hybrid representation makes ExAvatar fully

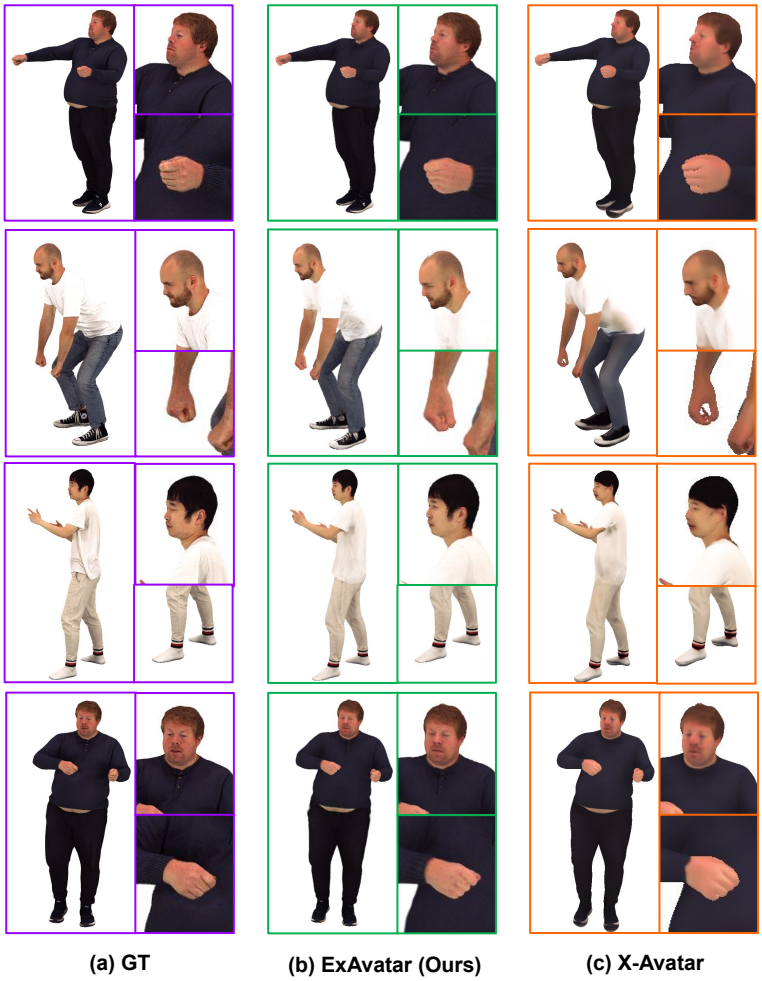

Fig. 9: Qualitative comparison between our ExAvatar and X-Avatar [41] on the test set of X-Humans [41].

compatible with the facial expression space of SMPL-X and significantly reduces artifacts in novel facial expressions and novel poses.

Limitations. First, as the inside of the mouth including the cavity and palm of the hands are often not observed in the video, our model hallucinates plausible geometry and textures. Second, like previous avatars [8,13,15–17,20,21,26,33,34, 41], ours struggles in modeling dynamic clothes. Material of clothes with motion information, such as velocity and acceleration, should be considered to properly model such dynamic clothes, out of our scope.

Future works. To hallucinate unobserved human parts better, such as inside of the mouth, score distillation sampling [35] can be used to *generate* images and use them for supervision. In addition, adding relightability to our ExAvatar is a promising and interesting future direction.

References

1. Alldieck, T., Magnor, M., Xu, W., Theobalt, C., Pons-Moll, G.: Video based reconstruction of 3d people models. In: CVPR (2018)
2. Alldieck, T., Xu, H., Sminchisescu, C.: imGHUM: Implicit generative models of 3D human shape and articulated pose. In: ICCV (2021)
3. Bagautdinov, T., Wu, C., Simon, T., Prada, F., Shiratori, T., Wei, S.E., Xu, W., Sheikh, Y., Saragih, J.: Driving-signal aware full-body avatars. ACM TOG (2021)
4. Cai, Z., Yin, W., Zeng, A., Wei, C., Sun, Q., Yanjun, W., Pang, H.E., Mei, H., Zhang, M., Zhang, L., et al.: SMPLer-X: Scaling up expressive human pose and shape estimation. NeurIPS (2023)
5. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., De Mello, S., Gallo, O., Guibas, L.J., Tremblay, J., Khamis, S., et al.: Efficient geometry-aware 3D generative adversarial networks. In: CVPR (2022)
6. Chen, J., Zhang, Y., Kang, D., Zhe, X., Bao, L., Jia, X., Lu, H.: Animatable neural radiance fields from monocular RGB videos. arXiv preprint arXiv:2106.13629 (2021)
7. Chen, Z., Moon, G., Guo, K., Cao, C., Pidhorskyi, S., Simon, T., Joshi, R., Dong, Y., Xu, Y., Pires, B., Wen, H., Evans, L., Peng, B., Buffalini, J., Trimble, A., McPhail, K., Schoeller, M., Yu, S.I., Romero, J., Zollhöfer, M., Sheikh, Y., Liu, Z., Saito, S.: URhand: Universal relightable hands. In: CVPR (2024)
8. Choi, H., Moon, G., Armando, M., Leroy, V., Lee, K.M., Rogez, G.: MonoNHR: Monocular neural human renderer. In: 3DV (2022)
9. Choutas, V., Pavlakos, G., Bolkart, T., Tzionas, D., Black, M.J.: Monocular expressive body regression through body-driven attention. In: ECCV (2020)
10. Contributors, M.: Openmmlab pose estimation toolbox and benchmark. `https://github.com/open-mmlab/mmpose` (2020)
11. Feng, Y., Choutas, V., Bolkart, T., Tzionas, D., Black, M.J.: Collaborative regression of expressive bodies using moderation. In: 3DV (2021)
12. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3D face model from in-the-wild images. ACM TOG (2021)
13. 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)
14. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: ICCV (2017)
15. 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. arXiv preprint arXiv:2312.02134 (2023)
16. Jiang, T., Chen, X., Song, J., Hilliges, O.: InstantAvatar: Learning avatars from monocular video in 60 seconds. In: CVPR (2023)
17. Jiang, W., Yi, K.M., Samei, G., Tuzel, O., Ranjan, A.: NeuMan: Neural human radiance field from a single video. In: ECCV (2022)
18. Joo, H., Simon, T., Sheikh, Y.: Total Capture: A 3D deformation model for tracking faces, hands, and bodies. In: CVPR (2018)
19. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3D gaussian splatting for real-time radiance field rendering. ACM TOG (2023)
20. Kocabas, M., Chang, J.H.R., Gabriel, J., Tuzel, O., Ranjan, A.: HUGS: Human gaussian splats. arXiv preprint arXiv:2311.17910 (2023)
21. Kwon, Y., Kim, D., Ceylan, D., Fuchs, H.: Neural Human Performer: Learning generalizable radiance fields for human performance rendering. NeurIPS (2021)

22. Li, J., Bian, S., Xu, C., Chen, Z., Yang, L., Lu, C.: HybrIK-X: Hybrid analytical-neural inverse kinematics for whole-body mesh recovery. arXiv preprint arXiv:2304.05690 (2023)
23. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4D scans. ACM TOG (2017)
24. Lin, J., Zeng, A., Wang, H., Zhang, L., Li, Y.: One-stage 3D whole-body mesh recovery with component aware transformer. In: CVPR (2023)
25. Liu, S., Li, T., Chen, W., Li, H.: Soft Rasterizer: A differentiable renderer for image-based 3D reasoning. In: ICCV (2019)
26. Liu, X., Wu, C., Liu, X., Liu, J., Wu, J., Zhao, C., Feng, H., Ding, E., Wang, J.: GEA: Reconstructing expressive 3D gaussian avatar from monocular video. arXiv preprint arXiv:2402.16607 (2024)
27. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM (2021)
28. Moon, G., Choi, H., Lee, K.M.: Accurate 3D hand pose estimation for whole-body 3D human mesh estimation. In: CVPRW (2022)
29. Moon, G., Shiratori, T., Lee, K.M.: DeepHandMesh: A weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling. In: ECCV (2020)
30. Moon, G., Xu, W., Joshi, R., Wu, C., Shiratori, T.: Authentic hand avatar from a phone scan via universal hand model. In: CVPR (2024)
31. Patel, P., Huang, C.H.P., Tesch, J., Hoffmann, D.T., Tripathi, S., Black, M.J.: AGORA: Avatars in geography optimized for regression analysis. In: CVPR (2021)
32. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: CVPR (2019)
33. 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)
34. 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)
35. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: DreamFusion: Text-to-3D using 2D diffusion. In: ICLR (2023)
36. Qian, Z., Wang, S., Mihajlovic, M., Geiger, A., Tang, S.: 3DGS-Avatar: Animatable avatars via deformable 3D gaussian splatting. In: CVPR (2024)
37. Ravi, N., Reizenstein, J., Novotny, D., Gordon, T., Lo, W.Y., Johnson, J., Gkioxari, G.: Accelerating 3D deep learning with pytorch3D. arXiv preprint arXiv:2007.08501 (2020)
38. Remelli, E., Bagautdinov, T., Saito, S., Wu, C., Simon, T., Wei, S.E., Guo, K., Cao, Z., Prada, F., Saragih, J., et al.: Drivable volumetric avatars using texel-aligned features. In: ACM SIGGRAPH Conference Proceedings (2022)
39. Rong, Y., Shiratori, T., Joo, H.: Frankmocap: A monocular 3D whole-body pose estimation system via regression and integration. In: ICCVW (2021)
40. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In: ICCV (2019)
41. Shen, K., Guo, C., Kaufmann, M., Zarate, J.J., Valentin, J., Song, J., Hilliges, O.: X-Avatar: Expressive human avatars. In: CVPR (2023)
42. 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)

43. Xu, H., Bazavan, E.G., Zanfir, A., Freeman, W.T., Sukthankar, R., Sminchisescu, C.: GHUM & GHUML: Generative 3D human shape and articulated pose models. In: CVPR (2020)
44. Zhang, H., Tian, Y., Zhang, Y., Li, M., An, L., Sun, Z., Liu, Y.: PyMAF-X: Towards well-aligned full-body model regression from monocular images. TPAMI (2023)
45. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)