GaussianAvatar: Towards Realistic Human Avatar Modeling from a Single Video via Animatable 3D Gaussians

Supplementary Material

In the supplementary material, we begin by presenting the implementation details of our method in Sec. 6. Following that, we provide information on the proposed dataset in Sec. 7, conduct the training & running time comparison in Sec. 8, and demonstrate the motion optimization comparison in Sec. 9. Finally, we showcase challenging cases in Sec. 10 and present hand animation results in Sec. 11.

6. Implementation Details

6.1. Model Architecture

We first estimate the SMPL model for all videos in three datasets. The input to the pose encoder is the UV map of the SMPL model, which has a resolution of 128 $\times$ 128 $\times$ 3. We adopt a standard U-Net architecture as the pose encoder, comprising five blocks of [Conv2d, BatchNorm, LeakyReLU], followed by five blocks of [ReLU, ConvTranspose2d, BatchNorm]. Note that we omit the BatchNorm in the final block.

The optimizable feature tensor has the same resolution as the output of the pose encoder, which is 128 $\times$ 128 $\times$ 64. During the first training stage, we train it using an auto-decoding approach. Subsequently, the output of the pose encoder is integrated into the optimized feature tensor before being input to the Gaussian parameter decoder. To achieve finer details, we conduct a $4\times$ upsampling of the combined feature tensor, resulting in a dimension of 512 $\times$ 512 $\times$ 64. The resulting output of 3D Gaussians consists of nearly 200,000 points.

The Gaussian parameter decoder comprises an 8-layer Multi-Layer Perceptron (MLP) followed by three prediction heads. The dimensions of the intermediate layers of the MLP are (128, 128, 128, 256, 128, 128, 128, 64), incorporating a skip connection from the input to the 4th layer. Each prediction head consists of a 2-layer MLP designed to predict offsets $\Delta\hat{\mathbf{x}}$, colors $\hat{\mathbf{c}}$, and scales $\hat{s}$, respectively.

6.2. Training

We first train the optimizable feature tensor and the Gaussian parameter decoder concurrently with motion optimization. During this stage, we employ the Adam optimizer with specific learning rates: 3.0×10^{-3} for the Gaussian parameter decoder, 5.0×10^{-4} for the optimizable feature tensor, and 5.0×10^{-3} for motion optimization. We train them for a duration of 200 epochs. Following this, we generate UV positional maps of SMPL models corrected by optimized motions. After the first stage of training, we suspend the training of the optimized feature tensor and combine it with the output of the pose encoder. We proceed to train the pose encoder and fine-tune the Gaussian parameter decoder for an additional 200 epochs.

Sequence	Total	Train	Validation	Test
male-1	978	782	98	98
female-1	972	778	97	97

Table 4. **Data distribution.** Number of frames in each sequence used for training, validation, and testing.

Methods	HumanNeRF	InstantAvatar	Ours
Training time	$\sim$ 13 h	$\sim$ **1 min**	$\sim$ 30 min
Running time	0.22 FPS	3.87 FPS	**35 FPS**

Table 5. **Training and running time comparisons.**

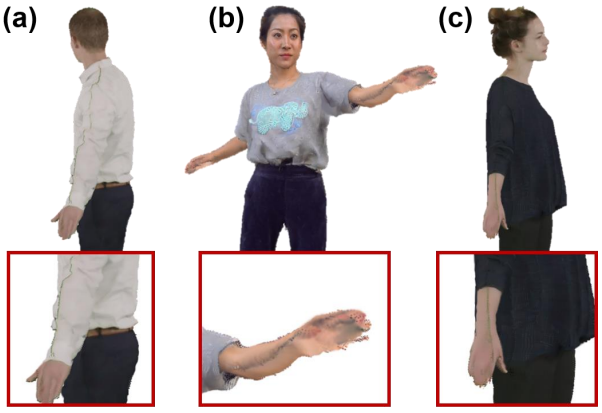

Figure 8. **Results of inaccurate segmentation.** We showcase the artifacts resulting from the inaccurate segmentation boundary.

7. Dataset Details

We take the same settings in NeuMan for partitioning the proposed DynVideo dataset. The dataset details are as shown in Table 4.

8. Training and Running Time Comparison

Here we compare the inference speed of GaussianAvatar with two NeRF-based methods, HumanNeRF and InstantAvatar. As shown in Table 5, we measure the training and running time in the People-Snapshot dataset.

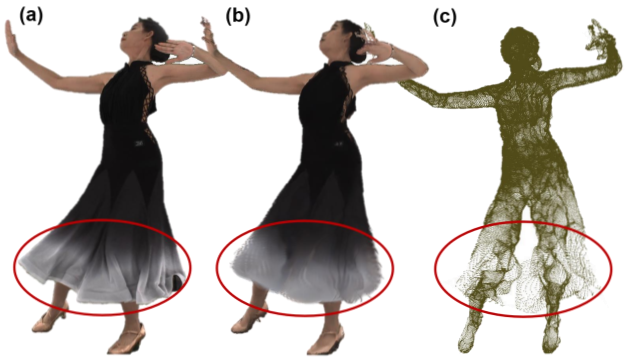

Figure 9. **Results of loose clothing.** (a) is the ground truth, (b) and (c) are the rendered image and Gaussian points.

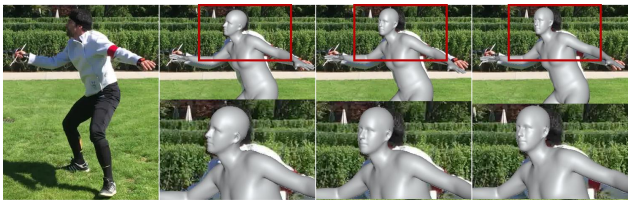

Figure 10. **Results of motion optimization comparison.** (a) Original image , (b) our optimized SMPL, (c) refined SMPL by InstantAvatar, (d) initial SMPL.

Methods	Initial motion	InstantAvatar	Ours
P-MPJPE	71.95	70.87	**64.94**

Table 6. **Motion optimization comparison.**

9. Motion Optimization Comparison

We directly evaluate the pose refinement of GaussianAvatar and the SOTA InstantAvatar on two sequences in the 3DPW dataset and one sequence in the DNA-Rendering dataset. Both Table 6 and Fig. 10 show that our GaussianAvatar outperforms InstantAvatar in pose refinement.

10. Challenging Cases

As discussed in the final section of the main paper, a major limitation of our approach is attributed to the inaccuracies in foreground segmentation in videos. As shown in Fig. 8, the inaccuracies in the foreground segmentation boundary may lead to our method predicting a black line on the surface. Automatic segmentation tools do not always yield satisfactory segmentation results. Manual operations on these segmentations are time-consuming and inefficient. We believe that addressing this issue can be achieved by incorporating a scene model, akin to approaches such as NeuMan and Vid2Avatar, which can contribute to more accurate segmentation. We leave this for future work.

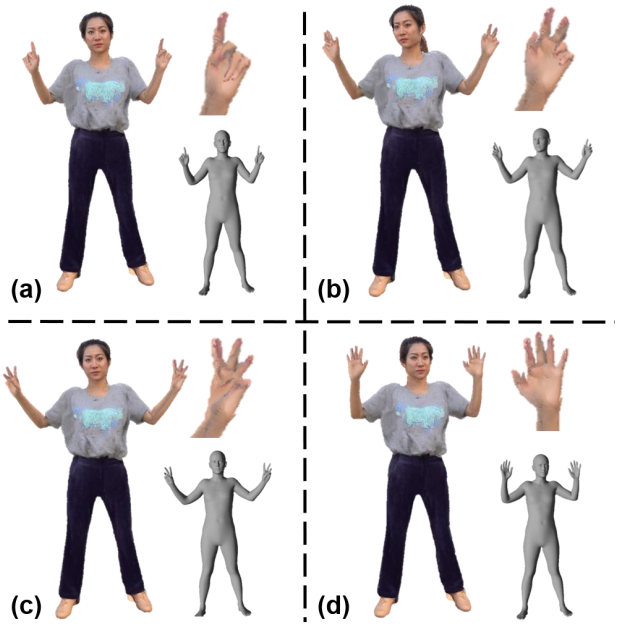

Figure 11. **Results of hand animation.** (a-d) Left: reposed image, bottom right: reference pose.

Besides, modeling the dynamic appearance of dresses remains challenging. As shown in Fig. 9, our method produces a blurred clothing appearance and fails to reconstruct complete point clouds. The primary challenge stems from the derived skinning weights from the SMPL model. Employing these skinning weights to model dresses may lead to artifacts when generalized to new poses. The prospect of predicting specific skinning weights for each subject is promising. However, this data-driven approach necessitates specific data sources. We intend to collect this kind of data in future efforts.

11. Hand Animation

We observe that our method can be readily extended to hand animation. To showcase its effectiveness in this context, we estimate the underlying SMPL-X model to fit a sequence from the DynVideo dataset. As depicted in Fig. 11, our method demonstrates the capability to generate plausible hand animation without the need for specific design considerations. The prospect of extending our work to encompass full-body avatars is promising, and we defer this to future investigations.