3D Gaussian Head Avatars with Expressive Dynamic Appearances by Compact Tensorial Representations

Supplementary Material

5. Implementation Details

Jaw Pose Linear Bases. We use farthest point sampling to extract linear jaw pose bases from jaw poses of each frames in videos, in order to unify dynamic textures due to linear blendshape and non-linear jaw rotation to linearly-interpolated feature lines. The jaw pose linear bases of id #074 are shown in Fig 7.

Class-balanced Sampling. The training frames are clustered into 16 categories and we evenly sample from 16 categories to ensure no bias toward expressions with less motion. We show the cluster center of subject #074 in Fig 8.

Acceleration. In our experiments, FLAME meshes are generated during the initialization stage to reduce time consumption during inference. Since dynamic textures are primarily concentrated on the face, the spatial bounds of feature lines are set around the face. Splats outside these bounds are excluded when calculating the opacity offset to further accelerate inference.

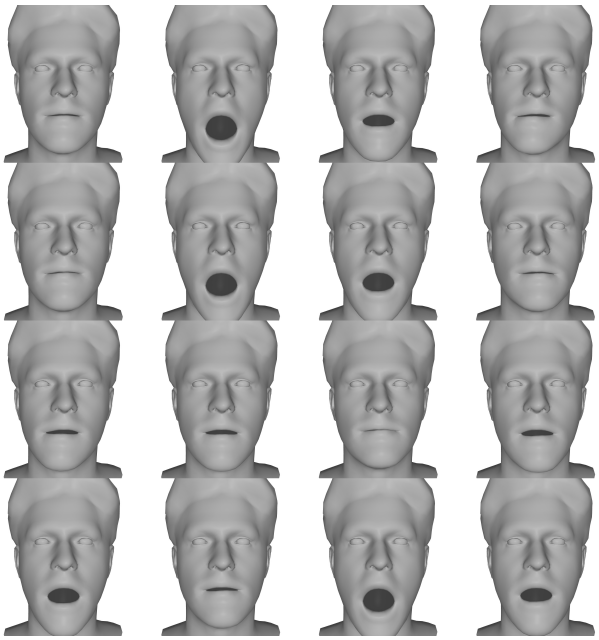

Figure 7. Basis jaw rotation extracted from all frames from videos via farthest point sampling of subject#074.

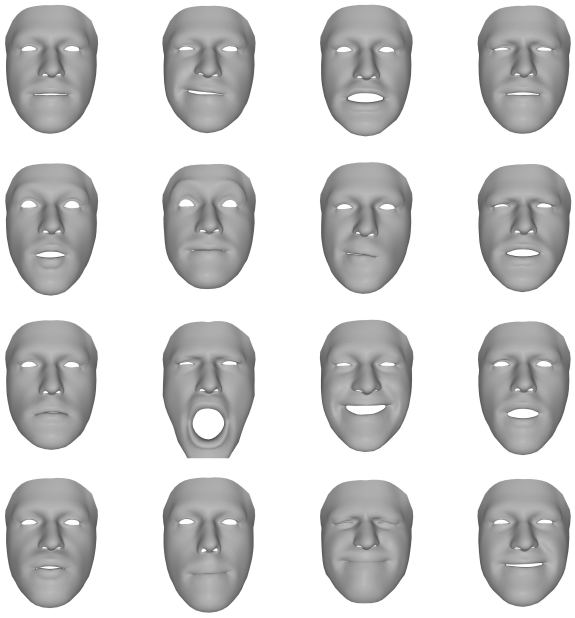

Figure 8. Cluster center for expression balanced sampling of subject#074.

6. Comparison Details

6.1. Baselines

We conduct comparative experiments with three baseline methods: GA, GHA, and GBS. To ensure fair comparisons, we align the inputs, including image resolution and pre-tracked mesh.

GHA. Both GHA and our method utilize multi-view videos from the Nersemble dataset, but the input resolutions differ. GHA processes 2K resolution images, while our input images are downsampled by a factor of four. To ensure fairness in testing, we first downsampled the 2K images by four times and then upsampled them back to their original size as the input for GHA. The FPS of GHA is tested by rendering 1024*1024 resolution images.

GBS. GBS is a monocular facial video reconstruction method, requiring monocular metrical tracker [44] to regress FLAME(2020 version, with two additional expression bases for describing closed eyes) coefficients and camera parameters from the images, which serve as the model's input. In our approach, the input consists of multi-view videos along with camera parameters and tracked FLAME(2023 version) coefficients.

Figure 9. Rendering results of extreme viewpoints and expressions. Extrapolated viewpoints are in the red box.

To ensure fairness of comparisons, we concatenate the multi-view videos into a single video and fit FLAME2020 coefficients to approximate FLAME2023 multi-view tracked meshes instead of monocular metrical tracker, serving as inputs of GBS. We optimize the FLAME 2020 parameters by calculating the mesh vertex positions loss. Note that the parameters output by the metrical tracker do not include the hair offset or neck motion, so we calculate the loss using only the facial region vertices. First, we compute the shape coefficients using the neutral expression, then regress the expression coefficients, eye rotation and jaw rotation for each frame in an iterative manner.

6.2. Dataset

We test our method on nine subjects (074, 104, 165, 175, 210, 218, 264, 302, 304) from Nersemble datasets. The free performace sequences are used to evaluate the effects of self-reenactment, which may contain some frames where the tongue is sticking out. As our method and compared baselines do not focus on mouth interior modeling, we exclude these frames from evaluation.

7. More Experiments

NeRF head avatar. INSTA [45] is a NeRF based head avatar method which enables fast training and inference. INSTA relies on FLAME mesh to guide NeRF to move correctly, which warps points according to the nearest mesh triangle directly.

	Novel View Synthesis			Self-Reenactment			Performance	
Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	Storage	FPS
INSTA	27.97	0.92	0.11	27.50	0.92	0.103	53M	25
Ours	32.97	0.95	0.059	28.07	0.93	0.077	10M	300

Table 3. Quantitative comparison with INSTA.

Offset on Opacity or Other Attributes. We tested adding offsets on position/rotation/scale of Gaussians to model face dynamics, but found this design leads to inferior performance of novel expression synthesis (Tab. 4 right), since it increases the model's degrees of freedom, making it prone to overfitting to training expressions, with weaker generalization to novel expressions.

	Novel View Synthesis			Self-Reenactment		
-	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Opacity	35.16	0.97	0.026	31.64	0.96	0.036
Others	36.39	0.97	0.018	30.97	0.96	0.030

Table 4. Ablation study of opacity offsets on subject #306. "Others" refers to position/rotation/scale offsets.

Extreme Viewports and Expressions. We conduct experiments to validate robustness of our method on extreme viewports and expressions. We interpolate 16 new turntable viewpoints from 16 training views, randomly select from 8 subjects and generate expression coefficients by sampling the first 8 dimensions of PCA space, which can be found in Figure 9.

8. Ethical Considerations

The generation of artificial portrait videos using our method poses risks, including the spread of false information, and erosion of trust in media credibility. These issues could have profound societal implications. Addressing this challenge requires developing reliable techniques to identify and verify authentic content.